FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Generative Adversarial Networks in Automated Chest Radiography Screening

Martim de Aguiar Quintas Penha e Sousa

DISSERTATION

Mestrado Integrado em Bioengenharia

Supervisor: João Manuel Pedrosa, PhD

Co-Supervisor: Ana Maria Mendonça, PhD

August 7, 2021

Generative Adversarial Networks in Automated Chest Radiography Screening

Martim de Aguiar Quintas Penha e Sousa

Mestrado Integrado em Bioengenharia

August 7, 2021

Abstract

Chest radiography (CXR) is one of the most common imaging modalities globally, playing an essential role in screening, diagnosis of several pathologies and disease management. However, CXR interpretation is a time-consuming and complex task, requiring the availability and resources of experienced radiologists. As such, automated diagnosis systems for multi-label pathology detection could play a major role in reducing the burden on radiologists and reduce the variability in image interpretation. Traditional machine learning and deep learning methods are usually applied for the diagnosis of specific pathologies, such as lung nodules, tuberculosis or, more recently, Covid-19. Modern computational capabilities allowed for the ascension of deep learning models and together with the increasing size of datasets, the development of multi-disease deep learning approaches has emerged, showing promising results.

Nevertheless, there are significant limitations in the developed algorithms. The lack of representative data and annotations can hinder the robust training of deep learning approaches. Despite the existence of large datasets, these tend to have highly unbalanced classes, with some pathologies being significantly more represented than other classes, which can lead to a degraded performance in the less-represented pathologies. Additionally, the presence of medical devices, annotations in the image or even the position of the patient can be interpreted by the algorithm as a proxy for certain pathologies, introducing bias, which is highly undesirable, as cases that do not fulfil these conditions will not be detected. Furthermore, deep learning models act as black-boxes, lacking the explainability of their decisions, which hinders the human understanding and the adoption of these methods in clinical practice.

Generative Adversarial Networks (GAN) could play a significant role as a solution for both of these challenges as they allow to artificially create new realistic images that are indistinguishable from the real ones. This way, new CXR images could be used to increase the prevalence of images in the less-representative pathologies, decrease the biases in the dataset and improve the explainability of the decisions by generating samples that serve as examples or counter-examples to the image being analysed, ensuring patient privacy.

The goal of this dissertation is to develop a GAN capable of generating high quality realistic artificial CXR images to tackle the limitations of data representation and decision explainability. To achieve this goal, a GAN variation, the Lightweight GAN, is trained on the VinDr-CXR dataset to generate high quality data, and then evaluated quantitatively and qualitatively, by a group of external evaluators of both radiologists and non-radiologists, and submitted to be used in two applications of image classification and pathology detection.

This work shows that by using the LWGAN is it possible to successfully generate realistic artificial CXR images with a small training dataset and reduced computational power. It is also shown that the generated images can improve training of pathology classification models, increasing the applicability in clinical scenarios of automatic CXR screening and diagnosis tools.

Keywords: Deep Learning, Generative Adversarial Networks, Chest Radiograph

Resumo

A radiografia torácica (CXR) é uma das modalidades de imagem mais comuns a nível mundial, desempenhando um papel essencial no rastreio, diagnóstico de várias patologias e gestão de doenças. Contudo, a interpretação de CXR é uma tarefa morosa e complexa, exigindo a disponibilidade e recursos de radiologistas experientes. Como tal, os sistemas automatizados de diagnóstico para a deteção de múltiplas patologias podem desempenhar um papel importante na redução da carga sobre os radiologistas e reduzir a variabilidade na interpretação de imagens. Os métodos tradicionais de *machine learning* e de *deep learning* são geralmente aplicados no diagnóstico de patologias específicas, tais como nódulos pulmonares, tuberculose, ou mais recentemente, Covid-19. As capacidades computacionais atuais permitiram a ascensão de modelos de *deep learning* e, juntamente com o aumento dos *datasets*, o desenvolvimento de abordagens de *deep learning* de múltiplas patologias tem vindo a emergir, mostrando resultados promissores.

No entanto, existem limitações significativas em relação aos algoritmos desenvolvidos. A falta de dados e anotações representativas pode dificultar o treino robusto de abordagens de *deep learning*. Apesar da existência de grandes datasets, estes tendem a ter classes altamente desequilibradas, sendo algumas patologias significativamente mais representadas do que outras classes, o que pode levar a um desempenho degradado nas patologias menos representadas. Adicionalmente, a presença de dispositivos médicos, anotações na imagem ou mesmo a posição do paciente pode ser interpretada pelo algoritmo como um *proxy* para certas patologias, introduzindo viés, o que é altamente indesejável, uma vez que não serão detetados casos que não cumpram estas condições. Além disso, os modelos de *deep learning* atuam como caixas negras, faltando a explicabilidade das suas decisões, o que dificulta a compreensão humana e a adoção destes métodos na prática clínica.

As *Generative Adversarial Networks* (GAN) poderiam desempenhar um papel significativo como solução para estes dois desafios, uma vez que permitem criar artificialmente novas imagens realistas que são indistinguíveis das imagens reais. Desta forma, novas imagens de CXR poderiam ser utilizadas para aumentar a prevalência de imagens com patologias menos representadas, diminuir os enviesamentos no dataset e melhorar a explicabilidade das decisões através da geração de amostras que sirvam de exemplo ou contra-exemplo à imagem a ser analisada, assegurando a privacidade dos pacientes.

O objetivo desta dissertação é desenvolver uma GAN capaz de gerar imagens CXR artificiais realistas e de alta qualidade para combater as limitações da representação de dados e da explicabilidade das decisões. Para atingir este objetivo, a variação da GAN, a *Lightweight GAN* (LWGAN), é treinada com o *dataset* VinDr-CXR para gerar dados de alta qualidade, sendo depois avaliada quantitativa e qualitativamente, por um grupo de avaliadores externos radiolgistas e não radiologistas, e submetida para ser utilizada em duas aplicações de classificação de imagem e deteção de patologias.

Este trabalho mostra que utilizando a LWGAN é possível gerar com sucesso imagens CXR artificiais realistas com um pequeno *dataset* de treino e capacidade de computação reduzida. Mostra-

se também que as imagens geradas conseguem melhorar o treino de modelos de classificação de patologias, aumentando a aplicabilidade em cenários clínicos de ferramentas de rastreio e diagnóstico de CXR.

Keywords: Aprendizagem Profunda, Redes Generativas Adversariais, Radiografia Torácica

Agradecimentos

Esta dissertação representa o culminar de um período de cinco anos na minha educação e desenvolvimento como pessoa. Considero este tempo como uma fase marcante na minha vida e não só pelos meus feitos académicos, mas também pelas relações, momentos e pessoas com quem tive a oportunidade de o passar.

Queria começar por agradecer ao meu supervisor e orientador João Pedrosa por toda a dedicação, disponibilidade, apoio e conhecimento que partilhou comigo nestes últimos nove meses. Por ter trabalhado neste projeto como se fosse dele e por me proporcionar a oportunidade de aprender muito durante este tempo. Considero que tive muita sorte por o ter tido como supervisor.

A todos aqueles que partilharam esta etapa comigo e que me ajudaram durante o desenvolvimento deste projeto no Laboratório de Imagem Biomédica do INESC TEC, obrigado. Queria agradecer especialmente à Professora Ana Maria Mendonça, Joana Rocha e Sofia Cardoso Pereira por toda a ajuda e apoio durante estes meses. Foram sem dúvida essenciais para esta dissertação e estou muito agradecido. Uma especial atenção ao Dr.Pedro e à Dr.Joana pela disponibilidade e conhecimento partilhado comigo.

Felizmente estou rodeado de pessoas boas que partilharam esta viagem comigo também fora do regime académico.

Quero começar por agradecer à minha família por todo o apoio e amor que nunca me faltou.

Aos meus amigos da faculdade, Rita, Inês, Filipe e Cristiana, obrigado pelos nossos momentos juntos e por terem tornado bioengenharia mais fácil, um especial obrigado à Inês por teres partilhado os últimos três anos desta fase comigo e teres sido a minha parceira, e também aos meus amigos desde muito antes da faculdade, que partilhemos mais momentos como aqueles até hoje.

À minha namorada Joaninha, que sempre me motivou a ser a melhor versão de mim próprio em todas as vertentes da minha vida, que me apoiou quando mais precisei e que me manteve com os pés assentes na terra. Obrigado por seres a minha melhor amiga.

Quero agradecer por fim à minha irmã, à minha mãe e ao meu pai por me terem apoiado incondicionalmente, por me terem proporcionado todas as oportunidades que fizeram de mim quem sou hoje. Não estaria aqui sem vocês e estou eternamente grato.

Martim Quintas e Sousa

"Opportunity is everything."

Anonymous

Contents

List of Figures

List of Tables

Abbreviations

ACGAN	Auxiliary Classifier Generative Adversarial Network
AdaIN	Adaptive Instance Normalization
AUC	Area Under the Curve
CAD	Computer Aided Diagnosis
CGAN	Conditional Generative Adversarial Network
CNN	Convolutional Neural Network
CT	Computed Tomography
CXR	Chest X-Ray
CycleGAN	Cycle Generative Adversarial Network
DCGAN	Deep Convolutional Generative Adversarial Network
DDR	Direct Digital Radiography
DR	Digital Radiography
EMA	Exponential Moving Average
FID	Fréchet Inception Distance
GAN	Generative Adversarial Network
GSA	Global Self-Attention
IDL	Interstitial Lung Disease
IS	Inception Score
IV	Intravenous
LWGAN	Lightweight GAN
LSGAN	Least Squares Generative Adversarial Network
MMD	Maximum Mean Discrepancy
MRI	Magnetic Resonance Imaging
MS	Mode Score
NIH	Nacional Institutes of Health
NLP	Natural Language Processing

PA	Postero-anterior
PET	Positron Emission Tomography
ProGAN	Progressive Generative Adversarial Network
ReLU	Rectified Linear Unit
ROC	Receiver Operating Characteristic
SE	Squeeze-and-Excitation
SFR	Screen Film Radiography
SLE	Skip-layer Excitation
SSIM	Structural Similarity Index
StyleGAN	Style Generative Adversarial Network
VTT	Visual Turing Test
wGAN	Wasserstein Generative Adversarial Network

Chapter 1

Introduction

1.1 Context

The thoracic cavity contains several major organs susceptible to trauma and disease. Respiratory and cardiovascular diseases are responsible for a large share of thoracic diseases. In 2015, 10.4 million people worldwide developed tuberculosis and 1.4 million died from it [1]. Lung cancer is responsible for killing 1.6 million people every year at an increasing rate [1]. Chest radiography is one of the most commonly performed medical examinations, with approximately 109 million CXR exams performed in Europe and a total of 3.6 million CXR exams performed in Portugal in 2015. Plain radiography is a very cost-effective method of screening and diagnosing pathologies and other findings. It is widely available and portable, however, with such an extensive amount of exams performed each year, radiologists suffer a burden due to the complexity and time-consuming characteristics of the task at hand.

CXR images show pathologies related to respiratory disease, such as lung opacities and pleural pathologies, and are widely used to show heart and vascular pathologies and medical devices. This shows the versatility of CXR images and suitability for diagnosing a large range of pathologies. One of the challenges, however, when interpreting CXR images is the complexity of the image, which can lead to variability between radiologists and incorrect diagnosis.

With this in mind, automated pathology screening and diagnosis systems have been rapidly increasing in popularity, with the aim of reliably identifying and characterizing pathologies in CXR images. The first automatic classification algorithms applied in the medical domain for CXR images were used for single-label classification, such as the detection of lung nodules [2]. With the increasing computational power and larger datasets, deep learning algorithms became more complex and multi-label pathology diagnosis systems emerged. Numerous deep learning methods have been proposed for multi-label classification of the most prevalent findings/pathologies [3–6].

1.2 Motivation

Publicly available medical information is a scarce resource due to privacy laws, high costs and slow bureaucratic processes. Initially, the available datasets were smaller due to the need of experienced radiologists for labeling, which is time consuming and expensive. Recent datasets are larger, but lack correctly annotated labels. Additionally, these datasets have highly unbalanced classes, having highly represented pathologies and less-represented pathologies, which can hinder the training of the deep learning models for the less-represented classes and result in an unoptimized performance. Besides the unbalanced classes, the datasets can introduce bias due to the presence of annotations, the positioning of the patients or even medical devices, such as pacemakers or other prosthetics, in the images. These factors can be interpreted by the algorithm as features of a certain pathology, which should be avoidable, as cases that do not have the same feature are not detected. Furthermore, deep learning algorithms have a black-box behaviour, lacking the explainability of their decisions, resulting in a hindering of human comprehension and the adoption of these methods in clinical practice.

To overcome these limitations, data augmentation techniques based on rotation, translation and cropping are used to amplify the scale of datasets, however, these samples are highly correlated with the existing ones and do not offer entirely new images to the training set. With the emergence of generative models such as GANs, new artificial high-quality data can be generated that is indistinguishable from real data. This way, less-represented pathologies could be complemented with newly generated artificial images, while decreasing bias. Furthermore, GANs can be used to give some degree of explainability by providing examples and counter-examples of the image being analysed without the need to resolve privacy concerns associated with the use of real images for this purpose.

Generative models such as GANs are, thus, a promising solution for this limitation and are growing in the medical context.

1.3 Goals

The goals of this project are:

- The development of a deep learning GAN for the generation of high-quality artificial CXR images resembling normal and abnormal images based on the images provided in the public datasets.
- The evaluation and clinical validation of the developed methods, which consist of mathematical evaluation metrics along with qualitative metrics, and visual evaluation by radiologists.
- The application of the developed methods for the generation of artificial CXR images in the context of training and explainability of a deep learning multi-label automated pathology detection system.

1.4 Contributions

This work aims at providing the following contributions:

- Development of a well performing GAN capable of generating high-quality CXR images, that can be used as a benchmark for future work.
- Provide a

1.5 Document structure

The remainder of this thesis is divided into seven Chapters. Chapter 2 focuses on Chest Radiography and the State of The Art of pathology detection. Chapter 3 presents the current approaches on Automatic Image Generation, along with the state-of-the-art evaluation metrics and optimization techniques, and the applications in Chest Radiography. Chapter 4 describes the methodology used for this thesis. In Chapter 5 the experimental work is described and detailed. Chapter 6 presents the results obtained throughout the development of this thesis. Chapter 7 focuses on the discussion of the results from the previous chapter. Finally, Chapter 8 aggregates the planned Future Work and the Conclusion of this thesis.

Chapter 2

Chest Radiography

In this chapter, an extensive review on general radiography is presented. This review includes the history of radiography, the physical principles behind it, the use and application of X-ray imaging, the state of the art of medical imaging and X-rays and a more detailed view of chest radiography.

2.1 Radiography

X-ray imaging is a non-invasive imaging diagnosis technique that works by emitting X-rays that pass through the human body. These are reflected and absorbed by the tissues and then captured by a film cassette. The use of X-rays for imaging techniques was discovered in 1895 by Prof. Wilhelm Conrad Röntgen. As he was experimenting with the newly discovered technology, he saw the bones of his hand on a photographic plate on the opposite side of an electron beam tube [7]. Afterwards, Röntgen also imaged the bones in his wife's hand, obtaining the world's first X-ray image. A few years later, portable X-ray machines were becoming a reality and were useful during the first world war.

2.1.1 Fundamental Physical Concepts

X-rays are a form of high energy radiation present in the universe. The wavelength (0,01 to 10 nanometers) of these rays is shorter than ultraviolet light but larger than gamma rays. X-rays were first discovered when experimenting with Crookes tubes, light bulb shaped tubes with a low pressure inside to minimize the amount of contained gas. These tubes are attached to a cathode and an anode, usually composed by a tungsten filament. When the cathode and the metal anode experience a large potential difference, electrons are accelerated in the vacuum, turning into X-rays. Most of the energy from the electrode is converted into thermal energy as it travels from the cathode and hits the anode, and the small amount that is not, is converted into the X-ray that will reflect from the 45 degree angle of the metal of the anode and travel to the object with which it collides.

An increase in the filament's temperature leads to an increase in the release of electrons and, therefore, an increase in the quantity of X-rays in the tube. As for the potential difference, an

increase in the voltage results in a higher kinetic energy and, therefore, a higher velocity, which will allow the X-ray to have more penetrating power.

The beam is received on a silver bromide plate sensitive to the electromagnetic radiation, leading to the production of black metallic silver from silver bromide. A comparatively small dose of X-rays is used to produce a subtle change in the plate, which is then amplified by chemical development to become visually identifiable [8].

Figure 2.1 shows an illustration of the working principle of an X-ray emitter.

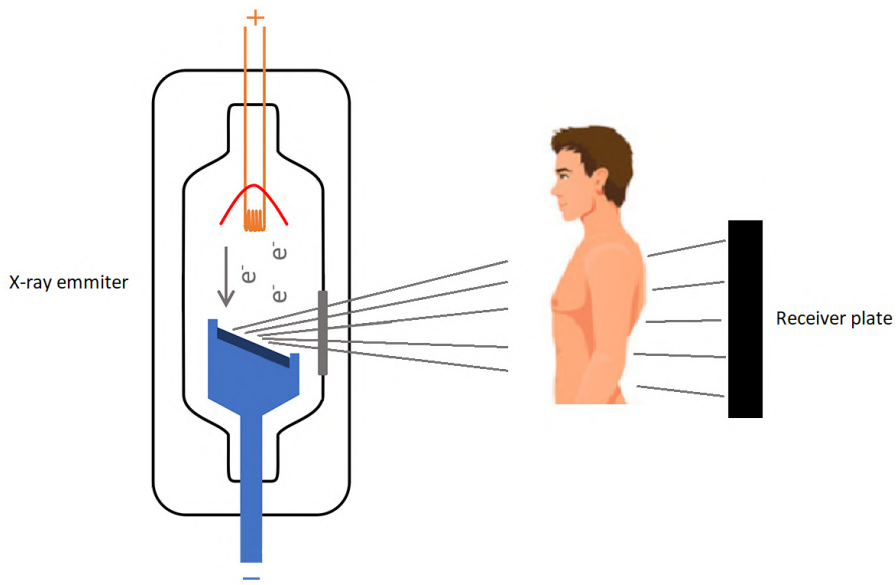


Figure 2.1: X-ray emitter along with the setup for collection of a CXR image. The potential difference between the cathode and the anode accelerates electrons that hit the metallic plate, producing x-rays. These go through the tissues and, with different energy levels, hit the receiver plate, creating an image.

The resulting X-rays finally collide with a film and change the color of that film. Depending on the anatomic structure that's being irradiated, each will retain different amounts of radiation, depending on the radiological density, resulting in absorption and reflection of different amounts of radiation. The end result of the collected X-rays is an image representing the shadows of each structure, where denser structures are colored in lighter tones, while less denser structures appear darker. Therefore, structures such as bones look white and structures as the lungs, due to their extensive air-filled volume, look darker. Figure 2.2 shows the color difference between bone and the tissue of a hand.

Figure 2.2: X-ray image of a hand [9]

2.1.2 Acquisition System and Image Formation

While this is how conventional X-ray works, it has now moved to digital. It was only in the 1980s that X-ray radiography became digital, and named computed radiography, by the use of a laser scanner to read the irradiated storage phosphor held in a cassette [10]. Today, digital radiography (DR) has become the main competitor against conventional screen film radiography (SFR).

As mentioned, SFR is the most common type of radiography used, however, there are certain limitations to this technology.

There are two main types of DR, computed radiography (CR) and direct radiography (DDR). The image acquisition, processing, storage, and display are four different steps with different processes. To maximize the efficiency of the whole process, the four sub-processes have to be optimized. CR uses a photostimulable phosphor plate for detection of X-rays, instead of the method used in SFR. A helium neon laser scans the exposed plate and a photomultiplier tube captures the emitted light. It is then converted to an analogue electrical system, which is later digitised. The other form of DR, DDR, cuts out the middle step of using latent image and an image plate reader by using a semiconductor-based sensor to directly convert X-ray energy into electrical signals. Solid state detectors and flat panel detectors are used as scintillators, which convert X-ray photons to light and later converting it to electrons through amorphous silica arranged as a photoiodide transistor [8]. Furthermore, DDR takes advantage of image intensification, which is used in real time images, by using a digital sensor linked to video monitors. This feature is very useful for screening several procedures such as vascular, orthopedic or radiological, due to its capability of increasing the brightness by up to 6000 times without an increase in the radiation dose.

Conventional SFR radiography is still the main technology used for plain radiography, however its lead to DR is progressively decreasing. The reasons behind this decreasing popularity are the fixed dose latitude, fixed non-linear grey scale response and the impossibility to decrease the radiation dose administered to the patient [8]. Other disadvantages include the lack of image processing and high cost of capturing the images, as with DR, the hospitals save money from lower film costs and lesser staff required to run the services. The film used is labour intensive, expensive and uses chemical processors, which contain hazardous materials, leading to higher costs when

compared to DR. Additionally, long term storage is difficult and the use of digital data archives is not compatible with SFR, in which the acquisition, processing and storage of the image are all done through the film. DR has the advantage of being available in digital format, which eliminates the need for storage and treatment rooms, saves time and allows for data exchange between clinicians and immediate access to archives. Additionally, DR is highly optimizable, as each of the four steps can be optimized separately. It is also faster, the imaging plates are reusable, the images do not deteriorate over time and can be processed for further analysis, avoiding another exposure to the patient.

2.2 Chest Radiography Imaging and Analysis

Plain radiography is a very cost-effective method of diagnosing and detecting pathologies and other findings. Besides, it is very accessible, as every hospital or medical clinic has radiographic equipment, and it can be portable. Portable X-ray machines provide care to patients in nursing homes, prisons and other facilities where taking the patient to a hospital is challenging [11]. Plain radiographs have also one of the best spatial resolutions (0.1mm) of all the imaging modalities.

X-ray images are used to diagnose and follow-up various types of conditions in a wide range of parts of the body. The most commonly analysed area is the chest with an average for 36 European countries of 194 images per 1000 of population every year, when compared to a frequency of 1100 X-ray examinations per 1000 of population for general X-ray procedures [12]. The number of CXR exams performed in 2015 in Portugal was around 3,6 million, while the European average for CXR exams was 109 million in the same year.

2.2.1 Other Chest Imaging Modalities

Given the wealth of clinical information that can be obtained from a single chest exam, there are naturally other imaging techniques that collect information in different ways, such as the computed tomography (CT), fluoroscopy, positron emission tomography (PET) and magnetic resonance imaging (MRI).

CT is an X-ray based non-invasive imaging technique that combines multiple X-ray images into a transverse two-dimensional view of the scanned area, which is used along with other parallel images to reconstruct a 3D volume CT image. The final result is a set of two-dimensional slices of a three-dimensional section [13]. Some of the advantages are its excellent depiction of anatomic detail, very fast exam time, possibility to examine organ enhancement as well as blood vessels with intravenous (IV) contrast and the three-dimensional view. However, it is relatively expensive, it may require IV contrast and it requires a higher dose of radiation when compared to plain radiography.

Fluoroscopy is a radiographic technique where several X-ray images are taken continuously to create an X-ray sequence with the use of a contrasting agent that is injected into the patient. The sequenced X-ray images allow the radiologist to have a detailed view of the path of the contrasting agent or the movement of a body part. This technique has its benefits, such as a detailed view

of several procedures in blood vessels, such as the implantation of stents, orthopedic surgery or barium X-rays to view the gastrointestinal tract. However, the patient can be exposed to X-rays for a long period of time for more complex procedures, resulting in a high dose of radiation being absorbed. Therefore, fluoroscopy procedures should be performed during a minimal time to avoid radiation risks [14].

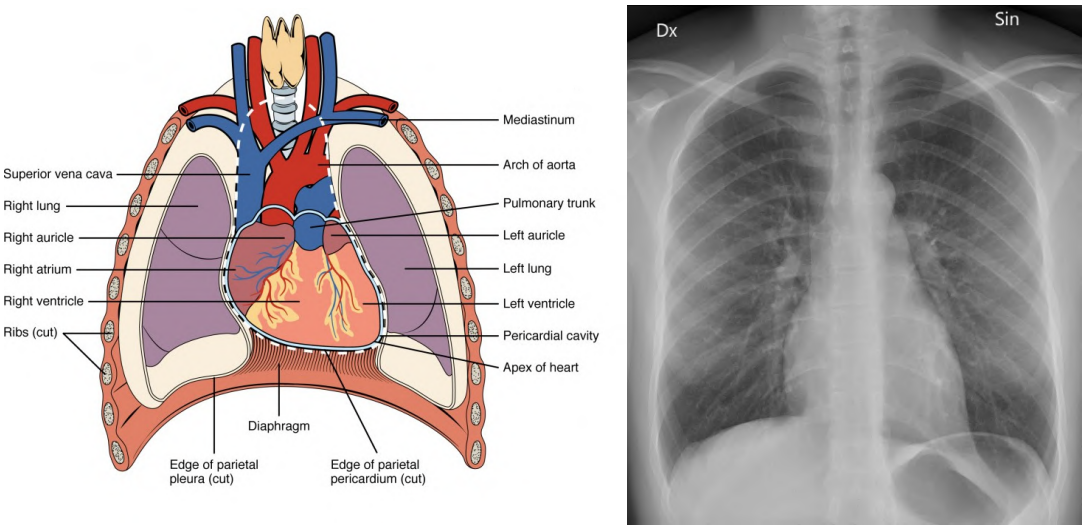
PET is a noninvasive imaging modality that provides physiological information through the injection of radioactive tracer compounds (radiotracers), detection of radiation, and reconstruction of the distribution of the radiotracer [15]. On one hand, this technology can provide information that other methodologies can not, such as increased activity in the cells by absorption of a larger quantity of radiotracers, indicating disease. Additionally, it can be combined with CT to provide further information and complement the other technique. On the other hand, there are fundamental trade-offs between resolution and noise and the quantitative accuracy of the measurements. Furthermore, PET requires a higher radiation dose when compared to the other conventional imaging techniques and is more expensive.

MRI is a non-invasive nuclear imaging technology that produces three-dimensional images of anatomical structures. The technology is based on the physical principle of "spin", exciting and detecting the change in the direction of the rotational axis of protons present in the water of living tissues. This technology is very expensive and is not widely available, with long waiting lists. Additionally, it is susceptible to the movement of the patient, due to its long examination period, and it requires preparation from the patient. Nevertheless, it allows for multiplanar and three-dimensional imaging, it produces high quality images, including images of soft tissue and organ contrast differences, and it requires no radiation [16].

When compared to these modalities, plain radiography has the advantage of requiring lower radiation doses, being widely available and cheap, and requiring very little preparation, making it ideal in screening and triage scenarios.

2.2.2 Chest Radiography in Clinical Practice

The thoracic cavity is enclosed by the spine, ribs and sternum, and the diaphragm separates the organs in the cavity from the abdomen. As shown in Figure 2.3a, it encloses the lungs, the heart, the aorta and esophagus among other organs. The pleura is a membrane that lines the whole extent of the thoracic cavity, covering the inside wall of the rib cage and spreads to the lungs. The purpose of the pleura is to lubricate the lungs by producing a liquid, protecting the lungs from friction during respiration [17].



(a) Anatomy of the chest cavity [18]

(b) CXR image

Figure 2.3: Thoracic cavity

Given the complex anatomical structures observable in a CXR image, a number of radiological findings can be observed by radiologists to infer a diagnosis. Some of the pathologies whose diagnosis is performed with or supported by CXR images are pneumonia, emphysema and cancer and it is through the analysis of these findings, together with other relevant clinical parameters (anamnesis, clinical reports, other exams, etc.) that can lead to a diagnosis. The same CXR image can, naturally, have multiple of these radiological findings. The main findings observable on CXR are described in Table 2.2, grouped into four categories related to the underlying anatomical structure where the findings are observed.

Heart and Greater Vessels

A classified enlarged mediastinum can result from findings such as the enlarged cardiomediastinum, cardiomegaly and a hilar prominence. An enlarged cardiomediastinum occurs mainly due to a cardiomegaly, which is an enlargement of the heart that occurs due to an increase in the transverse diameter of the cardiac silhouette, resulting in it being greater than or equal to 50% of the transverse diameter of the chest (increased cardiothoracic ratio) on a posterior-anterior projection of a chest radiograph or a computed tomography [19]. A hilar prominence is a unilateral or bilateral enlargement of the hila, which consist of vessels, bronchi and lymph nodes.

Lung Opacities

Findings related to lung opacity can result from infiltration and fibrosis (lung opacity), atelectasis, lung lesions (masses and nodules), pneumonia, consolidation and edema. These pathologies translate to the presence of denser material or structures, resulting in a higher level of opacity.

According to Barratt et al., pulmonary fibrosis is an interstitial lung disease characterised by chronic, progressive scarring of the lungs and the pathological hallmark of usual interstitial

pneumonia [20]. Pulmonary infiltration can have several infectious and non-infectious causes, such as acute leukemia, and it is defined as the presence of an abnormal substance that gradually accumulates in the tissue.

According to the National Heart, Lung and Blood Institute, lung atelectasis occurs when there's a collapse or incomplete expansion of pulmonary parenchima. This can happen due to conditions such as complete obstruction of an airway or due to a pneumothorax or pleural effusion, which limit the lungs' capacity to expand [21].

Masses and nodules are denser portions of mass that can be found in CT scans and in CXR images. Some of these nodules and masses may represent early disease and, importantly, can indicate lung cancer, which requires prompt diagnosis and definitive treatment.

Pneumonia is an infection caused by bacteria, virus of fungi. CXRs are commonly used to search for pneumonia because it is possible to distinguish it from other tract infections with this imaging modality [22].

Consolidation occurs when the alveoli and small airways in the lungs get filled with denser material. This can happen due to an infection, presence of blood, fluid or a cell mass.

CXR is also one of the most common methods to identify pulmonary edema. Pulmonary edema is defined by an abnormal accumulation of extravascular fluid in the lung parenchyma. This process leads to a diminished gas exchange at the alveolar level, progressing to potentially causing respiratory failure. It can be split into cardiogenic or noncardiogenic pulmonary edema [23].

Lung Pleura

Another set of findings relate to the pleura, such as pleural thickening, pneumothorax and pleural effusion. These pathologies relate to accumulation of air or fluid inside the pleura, which can result in a limited expansion of the lungs.

Pleural thickening occurs due to the presence of scar tissue in the lining of the lungs or the pleura. According to [24] it is the most common finding in CXR images.

A pleural effusion is an excessive accumulation of fluid in the pleural space, indicating an imbalance between the fluid production and removal by the pleura. This finding can have several causes, such as cancer, cirrhosis or tuberculosis [25].

Pneumothorax can be caused by physical trauma to the chest or as a complication of medical or surgical intervention (biopsy). Symptoms typically include chest pain and shortness of breath. Diagnosis of a pneumothorax requires a CXR or CT scan.

Other Findings

Other findings do not relate to one of the previous groups, such as fractures, medical devices, hernias and emphysemas. The identified medical devices are usually pacemakers or other prosthetics. As for hernias, a hiatal hernia is a condition in which parts of the abdominal contents, mainly the gastroesophageal junction and the stomach, are proximally displaced above the diaphragm through the esophageal hiatus into the mediastinum [26]. The other condition, pulmonary emphysema, is

a progressive lung disease in the form of chronic obstructive pulmonary disease, characterized by respiratory and airflow limitations usually caused by abnormalities, due to significant exposure to noxious particles or gases, such as the ones inhaled when smoking [27].

CXR images can have one or several pathologies, ranging from smaller areas of disease, such as masses or nodules, to larger areas such as the ones affected by an enlarged cardiomediastinum. Because of the extensive complexity of these images, radiologists spend large amounts of time analysing and labeling CXR images, still with uncertain results [28]. Additionally, as mentioned before, CXR is the most performed radiological exam worldwide, resulting in large numbers of exams to analyse and diagnose. These two factors combined result in a very time consuming and inefficient task. Thus, one of the solutions proposed for huge resource consumption is the implementation of automatic CXR pathology diagnostics systems [29].

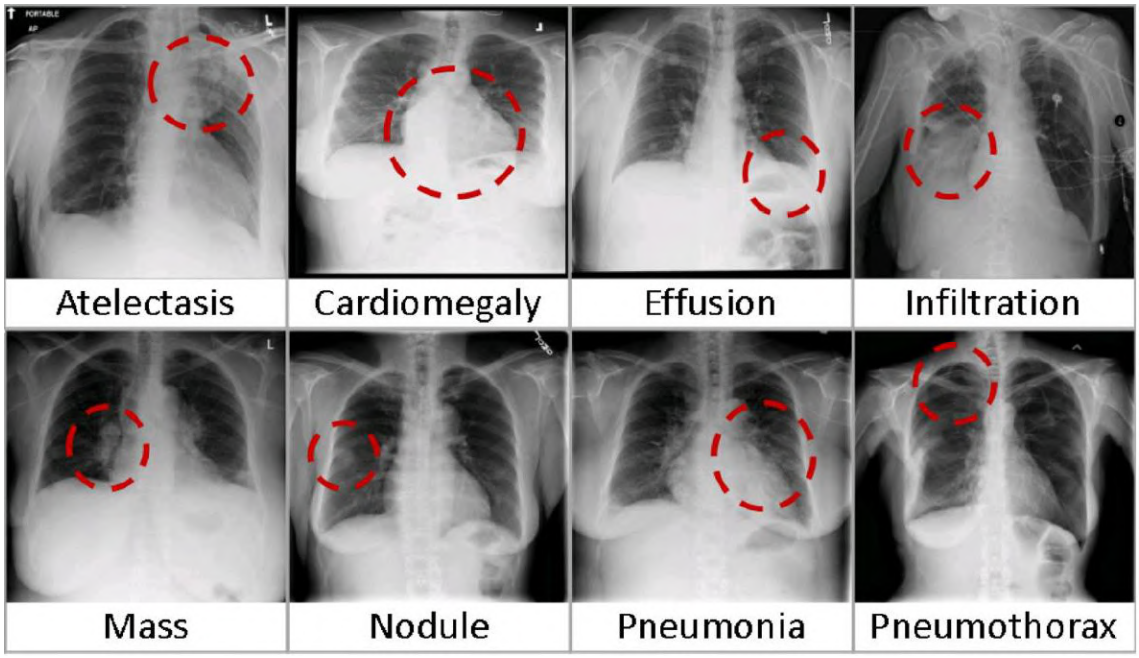


Figure 2.4: Eight common thorax pathologies found in the ChesX-ray8 dataset [30]

Figure 2.4 shows a set of commonly recurrent pathologies in the CXR images found in the one of the available public CXR datasets, ChestX-ray8. These images show the complex and, therefore, time consuming work performed by radiologists. As can be concluded, there is a need to develop automated systems capable of performing the classification of pathologies in these images.

2.3 Automatic Chest Radiography Analysis

Automatic computer aided systems for detection and diagnosis in CXR images can be based on traditional image analysis techniques, however, deep learning or machine learning methods have shown to be more accurate and have been evolving due to the advancement of new image signal

processing and more advanced methods. Traditional Computer Aided Diagnosis (CAD) systems rely on the manual extraction of visual features from image segments, so called rule-based processing as described by Bram van Ginneken [31]. Machine learning algorithms have the potential to be better alternatives against rule-based processing, however, feature extraction is still decided by the programmer and not by the computer. Deep learning differs from both methods in the way that it takes the images as inputs and puts them trough a network, passing them through multiple layers of processing steps and transforming them, having the feature extraction process occur in the intermediate layers of the network. Convolutional neural networks are considered the most promising path for the combination of feature extraction and classification in one [31].

Machine Learning

Machine Learning algorithms are systems that can learn and improve from experience without being explicitly programmed to do a task. These systems are based on a selection of features relative to the given data that are specified by the user, allowing the computer to learn and find patterns in the given data, resulting in the capability of making decisions based on that data.

Machine learning algorithms are divided into two fields: supervised learning and unsupervised learning. The former is a learning method based on labeled data, learning the input and the outcome and being able to learn from past examples in order to apply that knowledge to new unlabeled data using the labels of past data. The system is able to provide predictions for any new input after sufficient training. As for the latter, unsupervised learning methods receive data without labels, exploring it drawing inferences from datasets to describe hidden structures from unlabeled data [32].

In the medical research field, machine learning algorithms are used for image segmentation, abnormality detection and diagnosis of specific diseases, such as tuberculosis, lung nodules and Covid-19 [33]. In [34], an overview of the automatic machine learning based lung nodule detection systems was performed, collecting 41 studies related to the topic. Al-Timemy et al. propose a method based on machine learning for the detection of Covid-19 and tuberculosis in CXR images in undeveloped resource-limited countries [35].

Deep Learning

Machine learning algorithms fail to represent the wide array of pathologies encountered in the clinical environment. Increasing computational capabilities have led to the development of deep learning approaches with larger capabilities and better performances, making way for the development of multi-label detection approaches.

Deep Learning is a subset of machine learning with an increasing popularity for diagnostics systems and automatic pathology detection. It uses successive layers of representations to find hidden features and patterns in data. While in machine learning algorithms, the features are supplied by the programmer, in deep learning models the whole learning process is done through the use of

layered networks. Deep neural networks find the input-to-target pathway via a deep sequence of simple data transformations that are learned by exposure to examples [36].

Automatic CXR pathology diagnostics systems can be divided into two groups: abnormality detection and multi-label thoracic pathology classification systems. Abnormality detection systems classify CXRs as being normal or abnormal, where abnormal signifies that the patient is not healthy or at least one pathological finding was found. As for the multi-label classification systems, these identify the presence or absence of one or several pathological findings.

Abnormality detection systems are focused on more urgent abnormal cases, helping clinicians to manage their time. However, many CXRs have more than one pathology and single-label classification approaches are not adequate to evaluate these cases. With this problem in mind, multi-label classification approaches are being developed with the intent of covering many pathologies and not letting a major condition pass by unnoticed.

Regarding the abnormality detection approaches, Yates et al. proposed an algorithm based on Google's Inception convolutional neural networks (CNN), using transfer learning, such as the method proposed by [37], to overcome the computationally expensive task of training a CNN from scratch and training the binary normality classification of plain film CXR on the final layers of the neural networks [38]. For this approach, two datasets were used, the ChestX-ray14 database and the Indiana University hospital network chest radiograph (Open-I) database, having selected a mixed training set of frontal CXR images. In this approach, the authors did not use image augmentation, deeming it an unfit representation of the real-world dataset of CXR images. The end result of this approach showed promising results, having a final model accuracy of 94.6%, sensitivity of 94.6% and specificity of 93.4% with an Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) of 0.98.

Tang et al. presented a comparison between several already established CNN-based algorithms trained with the ChestX-ray14 database for binary image classification [3]. In this paper, the algorithms being compared are AlexNet, VGG16, VGG19, ResNet18, ResNet50, Inception-v3 and DenseNet121. All CNN's were both trained from scratch and also used transfer learning, pretraining with ImageNet, however the pretrained approaches outperformed the methods trained from scratch, having better AUC values. As for the comparison results, AlexNet achieved inferior results and DenseNet achieved the best results of all tested methods. The authors concluded that deeper neural networks do not significantly influence the performance of the binary classification task. Nevertheless, Dunnmon et al. concluded that while for larger datasets transfer learning methods or trained from scratch methods do not differ in accuracy, for moderate sized datasets this factor sets a significant difference in performance. The authors also presented a comparison between off-the-shelf methods, namely AlexNet, ResNet, and DenseNet. In this paper, ImageNet weights were used for pretraining. The final results showed a better performance for binary classification by the DenseNet algorithm [4].

Other approaches are based on alternative deep learning models, such as the approach proposed by Tuluptceva et al.. It consists of an autoencoder-based model, which values a subset of anomalous images to select the model's hyperparameters, on the contrary to common methods,

conveying a more flexible understanding of normality. Additionally, the authors chose to perform optimization with regard to perceptual loss in the regime of progressive growing training. Consequently, the deep perceptual autoencoder is capable of learning common patterns between normal observations and so accurately restore them, using the perceptual loss function to measure pattern dissimilarity [6]. In line with this method, Mao et al. propose a binary classification system based on autoencoders to take normal CXR images as inputs and simultaneously output the reconstructed input image along with the estimation of the pixel-wise uncertainty in reconstruction [5]. The reconstruction of normal images is prone to relatively large reconstruction errors around the boundaries of different regions, resulting in false abnormal classifications. This way, abnormal images can be identified considering the uncertainty-weighted reconstruction error as a measurement for abnormality presence.

As these methods evolve, different approaches emerge. In the case of the work developed by Tang et al., the authors propose an architecture for abnormal chest X-ray identification using generative adversarial one-class learning. In this approach, three deep neural networks are trained with normal CXR images while competing and cooperating to better model the underlying structure of normal images, resulting in a model fitted to properly reconstruct normal images and poorly reconstruct abnormal images [3].

As mentioned, several classification systems try to identify one or several pathologies in CXR images, the multi-label classification systems.

Guan et al. proposed a multi-label pathology classification approach based on a three-branch attention guided CNN [39], similar to the one proposed by [30]. Since some of the most common pathologies lie in a smaller area of the CXR images, the noise of the irrelevant part of the image can significantly impact the learning of the model. However, focusing only on the disease area can lead to information loss. The solution proposed by the authors consists of a global and a local branch classification networks, complementary to each other, where the global branch is trained with the global images and the local branch with the cropped disease area, being both fused together in a final fine-tuning network. In a similar work, Chen et al. propose a classification method based on two asymmetric subnetworks, a ResNet and a DenseNet, that complement each other with the asymmetric features collected by each network [40]. Each network learns different features from several pathologies. Both asymmetric feature sets are later fused in a third classifier, preserving the complementary discriminative features, and being evaluated by a unified loss function, speeding up the convergence of the training phase. The authors also support an iterative training strategy, optimizing each asymmetric stream alternatively, effectively improving the generalization ability of the proposed approach.

Lastly, Zhang et al. introduce a classification algorithm based on the use of weakly supervised distance learning to learn discriminative features from triplets of images and region verification module that feeds back class-specific common attentive regions [41]. The purpose of using distance learning is to set images with common pathologies in a nearer feature space. As for the region verification, the authors propose to feed another region classifier with the common attentive regions.

The methods mentioned above show the capabilities of deep learning algorithms in multi-label classification problems and show the progressive growth in the variety of these method in the search for the best solution.

2.3.1 Datasets

Given the dependence of deep learning in large amounts of data, an effort has been made by several entities to create larger and higher quality CXR datasets. These enable the improvement of the overall deep learning models' performance, and a higher level of state-of-the-art methods for automated diagnosis systems.

	Nº of findings	**Nº of samples**	**Image label**	**Label Method**
ChestX-ray8 [30]	8	108 948	Global and Local	NLP
ChestX-ray14	14	112 120	Global	NLP
CheXpert [42]	14	224 316	Global	NLP
MIMIC-CXR [43]	14	377 110	Global	NLP
VinDr-CXR [44]	28	18 000	Local	Manual

Table 2.1: List of the five largest CXR datasets with corresponding number of images, labels and labelling method

One of the most used dataset is the ChestX-ray8. This dataset comprises 108 948 front-view CXR images collected between 1992 and 2015 from 32 717 different patients, as shown in Table 2.1. Each image is labeled with one or several of the eight most common disease labels. The images were diagnosed and classified by radiologists, with some of them containing local labels. The corresponding labels for the images were mined using Natural Language Processing (NLP) algorithms from the medical reports [30].

Later on, the ChestX-ray8 evolved into the ChestX-ray14. The improved dataset became the largest and highest quality dataset available in 2017, with 112 120 images from 30 805 unique patients, labeling 14 instead of 8 common pathologies. This dataset was created by the National Institutes of Health (NIH) and is widely used by the state-of-the-art methods for automatic pathology diagnosis systems [45].

Developed in 2019, the MIMIC-CXR dataset is the largest CXR images dataset, with 377 110 front-view and lateral-view images taken from 65 379 patients between 2011 and 2016 from the Beth Israel Deaconess Medical Center. Each image was diagnosed and classified by a radiologist. The extraction of the labels for each image also derived from two NLP algorithms applied to the radiology reports associated to each image [43, 46].

Besides the datasets mentioned, there is also the CheXpert dataset. This public dataset labels the images with one or several of 14 common pathologies, similar to the ones used in the ChestX-ray14 dataset, however there are additional labels for medical devices and fractures. The dataset is comprised by 224 316 front-view and side-view CXR images collected from 65 240 patients

between 2002 and 2017 by the Stanford Hospital [42]. The labeling for the CheXpert images is done using an automated rule-based labeler.

Lastly, the VinDr-CXR dataset is the most recent collection of images available. It contains more than 100 thousand CXR images. Nevertheless, this dataset is still unlabeled for the majority of the images. In December 2020, a set of 18 000 images manually labeled by radiologists was released to the public, becoming the largest dataset to be manually labeled, containing bounding boxes limiting the region of the identified finding [44].

The identified labels of each dataset are shown in Table 2.2.

2.4 Towards Robust Chest Radiography Pathology Detection

The search for highly accurate automatic CXR pathology detection systems has proven to show promising results, as mentioned in the related work. Nevertheless, these systems still pose some limitations regarding the models and the datasets. The limitations seen in the developed models are taken into account in the modern developments for promising models, however, the limitations with the datasets are harder to solve, due to the fact that the datasets are created by a singular entity.

The current available datasets provide large amounts of high-quality information to be used in automatic CXR pathology detection systems, however, there are still several barriers to be overcome. The datasets mentioned above are limited by a list of predefined labels without the location of the findings, lacking the ground truth location for the abnormalities. Some datasets, such as the VinDr and the ChestX-ray8 datasets, have a very limited set of images with bounding boxes to locate the findings, but with a high inter-observer variability between the radiologists.

Another major problem identified by many authors is the accuracy of the NLP algorithms for the label retrieval [44]. The NLP algorithms and the rule-based labelers are inconsistent, uncertain and can deliver wrong labels, still being unclear on how to annotate the large amounts of images, which can result in a poor performance of the deep learning models [47,48]. Although the larger datasets provide large amounts of images, the labels are not accurate in some cases. As for the datasets that have manually labeled images, most still lack the size of the larger datasets, however, the VinDr dataset is an example of an acceptably large dataset manually labeled on local regions [49].

As stated by [40], having global images as inputs to train new deep learning models leads to a lot of noise-related features, due to the complicated background. The excess noise can generate distractions and misguide the learning phase of the models. Additionally, noisy backgrounds hinder the task of detecting smaller abnormalities, such as masses or nodules.

Another noise-related problem mentioned by Guan et al. is the existence of irregular borders in the CXR images, due to poor alignment of the patient. This can result in a significantly negative effect on the classification accuracy [39].

The major problem that still cannot be solved by an optimized labeling algorithm is the class imbalance. Some pathologies are much rarer than others, resulting in an insufficient amount of

	ChestX-ray8	ChestX-ray14	CheXpert	MIMIC-CXR	VinDr-CXR
Heart and Greater Vessels			Enlarged Mediastinum	Enlarged Mediastinum	
					Aortic Enlargement
	Cardiomegaly	Cardiomegaly	Cardiomegaly	Cardiomegaly	Cardiomegaly
	Hilar Prominence				
Lung Opacity			Lung Opacity	Lung Opacity	Lung Opacity
	Infiltration	Infiltration			Infiltration
					Calcification
		Fibrosis			Pulmonary Fibrosis
					Interstitial Lung Disease
	Atelectasis	Atelectasis	Atelectasis	Atelectasis	
	Mass	Mass	Lung Lesion	Lung Lesion	Nodule/Mass
	Nodule	Nodule			
	Pneumonia	Pneumonia	Pneumonia	Pneumonia	Consolidation
		Consolidation	Consolidation	Consolidation	
		Edema	Edema	Edema	
Pleural		Pleural Thickening	Pleural Other	Pleural Other	Pleural Thickening
	Effusion	Effusion	Effusion	Effusion	Pleural Effusion
	Pneumothorax	Pneumothorax	Pneumothorax	Pneumothorax	Pneumothorax
Other Findings			Fracture	Fracture	Other Lesion
			Support Devices	Support Devices	
		Hernia			
		Emphysema			

Table 2.2: List of labels in each dataset

images to be used for training the deep learning algorithms. This uneven division leads to poor performance in the detection of some pathologies. Parallel to this notion of the lack of information on the location of the findings, is the lack of explainability. Deep learning algorithms have a black-box nature, which limits the clinical use of developed algorithms. Explainability is defined as a set of domain features such as pixels of an image that contribute to the output decision of the model. These models lack the ability to explicitly represent the knowledge for a given performed task. This limitation fails to assure the transparency and fails to gain the trust of the medical community [50].

Additionally, these algorithms find patterns, such as annotations or the positioning of the patient, that are not based on legitimate features, leading to a biased classification. These biased decisions can affect the whole database or just some classes or features. For this reason, it is necessary to have some degree of explainability to assure the correct unbiased decision by the system.

Regarding the poor alignment of the CXR images, despite it being a setup difficulty and not a technical one, some solutions include discovering the lesion region by mapping or dividing the image in separate segments and training each segment individually.

Generative models are able to artificially create new CXR images based on real normal and abnormal images. These images can be used for training, increasing the amount of minority class images containing rare pathologies and decreasing biases present in the datasets, while avoiding the privacy concerns associated with the use of real images in clinical testing. Roth et al. demonstrate the end result of applying data augmentation strategies by comparing the same models with and without it [51]. Furthermore, these models can provide some explainability of the decisions made by the deep learning solutions, by generating examples and counterexamples, which have strong similarities and lack the similarities, respectively, to the images being analysed. This feature is of great importance to understand the decision for each image.

With all of the above arguments in mind, the proposition of this dissertation is to use generative models to create new CXR examples to balance the uneven distribution of the abnormality classes, provide explainability for each decision, decrease bias in the dataset, eliminate heterogeneous setup and appearance of CXR across datasets, and to create a method of obtaining new realistic images for data augmentation.

Chapter 3

Automatic Image Generation in Chest Radiography

In this chapter, the state-of-the-art of the generative models will be presented. Additionally, the applications of generative models in CXR images will also be described.

3.1 State of the Art in Image Generation

Given its potential applications, artificial image generation has been the subject of extensive research and literature. Besides what is discussed in Section 2.4, image generation could be used for image synthesis, reconstruction, segmentation, noise reduction and classification purposes [52].

Given a set of images, generative models aim at artificially generating new images by learning the distribution of the training data. Based on unsupervised learning, generative models generate data from a vector of random numbers, called latent space. However, some models can generate images from other images. The learning phase of a generative model assures that the model creates a correct sample based on the features of the training set. The generative process of data retains value due to the fact that it naturally expresses casual relations of the context of the data, instead of just generalizing from mere correlations [53].

The most commonly used generative models are Variational Autoencoders (VAE) and Generative Adversarial Networks (GAN). While for GANs the focus is on how to arrive at a model that approximates the input distribution, VAEs attempt to model the input distribution from a decodable continuous latent space [54]. This is one of the reasons why GANs are able to generate more realistic data when compared to VAEs. GANs produce high-resolution and realistic-looking images, while VAEs generate images that are less sharp. The major disadvantage of GANs is the proven difficulty of efficiently and correctly training the models. However, with the correct stable training, GANs have a powerful capability of generating new data that cannot be met by a VAE or other generative models. For these reasons, the state of the art of this section is limited to GANs and these will be the main tools used for the generation of high quality and resolution CXR images in this dissertation.

Generative Adversarial Networks were first proposed in 2014 by Ian Goodfellow [55]. A GAN is an unsupervised learning method, meaning that it can be trained with unlabeled data and learn density distribution of data, the internal representations, generating samples that are closely similar to real data.

GANs are capable of generating artificially realistic sets of data such as images or sounds. The particular characteristic about GANs is the adversarial component of the network. This algorithm learns through a process of iteration between a generator (G) and a discriminator (D). When used to artificially generate images, G iteratively learns to create better quality images, both at the pixel-level details and at the larger scale features. D, on the other hand, is a classifier network whose purpose is to classify the image created by G as real of fake.

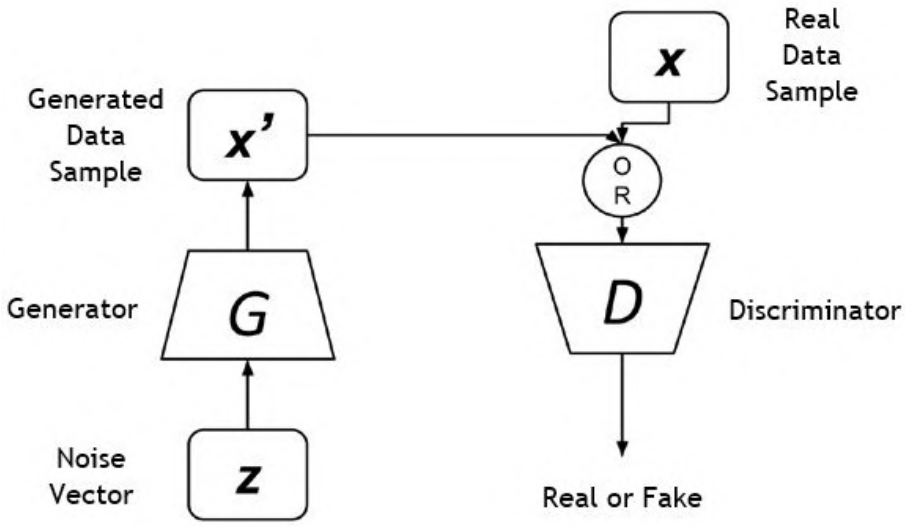


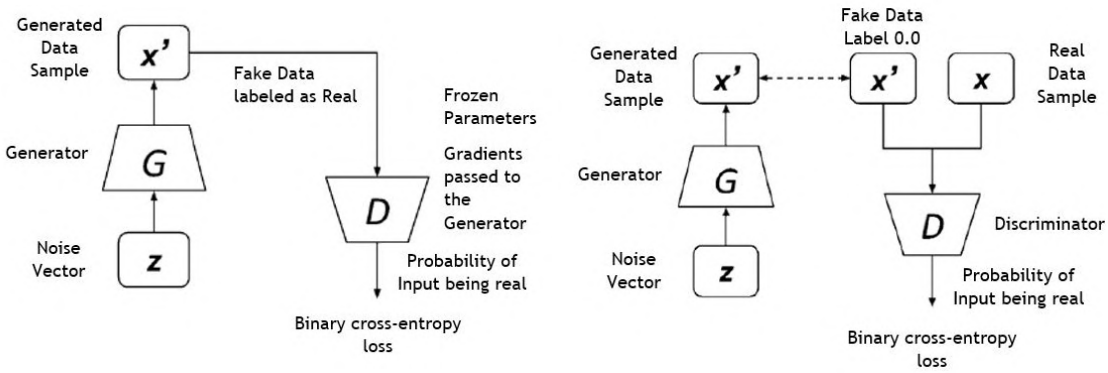
Figure 3.1: Simple GAN representation: Generator and Discriminator [54]

G captures the data distribution and D estimates the probability of a sample coming from the model distribution or the data distribution [55]. The rationale behind the training of G is to maximize the probability of the classification by D being a mistake.

One of the most commonly used comparisons for GANs is the one of a counterfeiter artist and an art expert. G, the counterfeiter, tries to replicate a certain painting style from an artist by learning from completed artwork. D, the art expert, classifies the forged artwork as looking real or fake, and both G and D learn from this feedback, iteratively improving the quality of the paintings created until D can no longer distinguish a real from a fake.

The process consists of the input of a noise vector into the generator network, resulting in an artificially generated image. Meanwhile, the discriminator receives a generated sample or a true sampled data. The discriminator network labels the valid or real data with 1.0 and the fake data with 0.0, meaning a 100% probability of it being sampled from the real data and 0% probability of it being real, respectively. The discriminator network learns from the supplied dataset on how to distinguish real data samples from generated ones and its parameters are updated with the input of real and fake data. Occasionally, the generator network creates an artificial batch of samples that is

fed to the discriminator network as real data and the discriminator should classify it as fake. At this time, the GAN will let the backpropagation of the gradients from the last layer of the discriminator network to the first layer of the generator network. However, during this iteration, the discriminator parameters are temporarily frozen. The goal of the process for the generator is to iteratively learn from the feedback of the discriminator network, improving gradually. Once the discriminator network can't distinguish the generated sample from real sampled data, the generator's learning phase is complete and the discriminator is discarded [54]. At this point the generator network is capable of generating artificial samples with a realistic look and high quality.



(a) Training process of the Generator network

(b) Training process of the Discriminator network

Figure 3.2: Training phase of the two networks that constitute a GAN [54]

The training process of the overall model is often described as a zero-sum game, known as min-max game. The discriminator network can be trained by minimizing the loss function described by Equation 3.1. The loss function is a standard binary cross-entropy cost function. The training cost is evaluated based on two differentiable functions, one for each network, where the loss is defined by the negative sum of the expectation of successfully identifying real data, *D(x)*, and that same expectation of 1.0 minus correctly identifying the synthetic data, *1-D(G(x))*. It is described by the loss function L, that depends on both networks. Real data labeled with 1.0 is defined as $x \sim p_{data}$ and z is the noise vector used by the generator.

$$L(G,D) = \mathbb{E}_{x \sim p_{data}} logD(x) + \mathbb{E}_z log(1 - D(G(z))) \tag{3.1}$$

To minimize the loss function, the discriminator parameters, represented by D, are updated by backpropagation, by correctly classifying the real data, *D(x)*, and artificial data, *1-D(G(z))*, where correctly identified real data translates to *D(x)* being close to 1.0 and correctly identifying fake data is equivalent to *D(G(z))* being close to 0.0.

Due to the loss function being a zero-sum game, the generator loss function is the negative of the discriminator loss function:

$$L^{(G)}(G,D) = -L^{(D)}(G,D) \tag{3.2}$$

converted to a value function:

$$V^{(G)}(G,D) = -L^{(D)}(G,D) \tag{3.3}$$

According to Equation 3.3, the equation should be minimized, from the perspective of the generator, and maximized, from the discriminator's point of view.

$$G^* = \min_G \max_D V^{(G)}(G,D) \tag{3.4}$$

Periodically, the generator tries to fool the discriminator by feeding a fake data sample with a label of 1.0. By maximizing with respect to D, gradient updates are sent to the discriminator parameters to consider the fake data as real. Simultaneously, by minimizing with respect to G, the generators parameters are trained on how to fool the discriminator network. The gradient updates are small and increasingly diminished as they propagate through the generator's layers, sometimes leading to non-convergence, which consists of both networks' parameters failing to stabilize and converge. The loss function of the generator results in:

$$L^{(G)}(G,D) = -\mathbb{E}_z \, logD(G(z)) \tag{3.5}$$

This function maximizes the probability of the discriminator network believing the artificial data is real, by training the generator.

3.1.1 Deep Convolutional GAN

GANs are known for being unstable at training, which often results in a generator that produces wrong outputs, and are affected by several factors:

- mode collapse, defined by the generation of different samples from the latent space to the same output, resulting in similarly generated images,
- diminished gradients, where the discriminator network learns at a fast pace and becomes successful, resulting in a vanishing gradient in the generator network,
- a stage of non-convergence, where the generator and the discriminator fail to stabilize and converge,
- high sensitivity to hyper-parameters.

The Deep Convolutional GAN (DCGAN), firstly proposed by Radford et al., is a CNN based GAN that became a success at improving the training stability of GANs [56]. The most notable change regarding this method is the swap of the maxpoolings with strided convolutions in the discriminator network and fractionally strided convolutions in the generator network. This allowed the CNNs to resize the feature maps. Apart from this change, several other aspects are characteristic from the DCGAN such as:

- Avoiding Dense layers, replacing with CNN in all layers, except the first layer of the generator, for it accepts the z-vector
- The batch normalization, used for stabilizing learning by normalizing each layer's input to have zero mean and unit variance
- Rectified Linear Unit (ReLU) is used in all layers of the generator network, except in the output layer, which uses Tanh
- Leaky ReLU in all layers of the discriminator [54]
- Adam optimizer instead of Stochastic Gradient Descent with momentum [57]

.

3.1.2 Conditional GAN

With the DCGAN there is no control over the specific output to be produced by the generator. There is no mechanism for requesting a specific output, therefore the generated images are random and cannot be used to fulfill a specific purpose or request.

Conditional GANs (CGAN) are based on DCGANs. A CGAN differentiates itself from the original architecture by the addition of an imposed condition label to the generator and the discriminator. CGAN is similar to DCGAN in most of its extent except for the additional one-hot vector input. For the generator network, this label is concatenated with the first layer, and with the discriminator network, a new layer is added. This layer has the sole purpose of processing the one-hot vector for the discriminator. The fundamental working principle of a CGAN is the same as the original GAN, however, the generator and discriminator networks are conditioned on one-hot labels.

An example of the application of the CGAN is the generation of digits, such as the developed work with the MNIST dataset [58], where a CGAN would generate specific digits conditioned by the one-hot labels [54]. Another use of the CGAN in the medical domain is described in [59] for the task of lung segmentation on a given CXR. The CGAN was trained to generate a segmented mask of a given input CXR, while being as realistic as possible compared to the ground truth masks.

Taking into account the incorporation of the condition, the resulting loss function for the discriminator and generator are shown in Equation 3.6 and 3.7, respectively:

$$\max_{D}\min_{G} V^{(D)}(G,D) = \mathbb{E}_{x\sim p_{data}(x)}\, logD(x|y) + \mathbb{E}_{z\sim p_z(Z)}\, log(1-D(G(z|y))) \tag{3.6}$$

$$\max_{D}\min_{G} V^{(G)}(G,D) = \mathbb{E}_{z}\, logD(G(z|y)) \tag{3.7}$$

The improved loss function from the discriminator network aims at minimizing the error of falsely predicting the true label of images, real images sampled from the dataset and fake generated images, given their one-hot label.

Figure 3.3 shows an illustration of the training phase of a CGAN's discriminator network, where both the generated and real images are conditioned with their corresponding one-hot label.

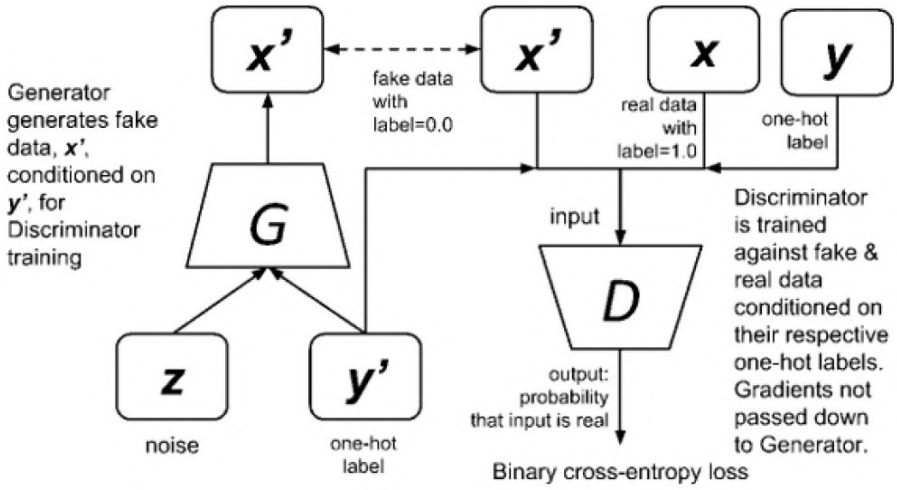


Figure 3.3: Training of the CGAN discriminator, where both the generated and real images are conditioned by a one-hot label [54]

As for the training phase of the generator, shown in Figure 3.4, it aims at minimizing the correct prediction of the discriminator on fake one-hot labeled conditioned images. The generator network learns how to generate a specific image according to its one-hot vector, fooling the discriminator.

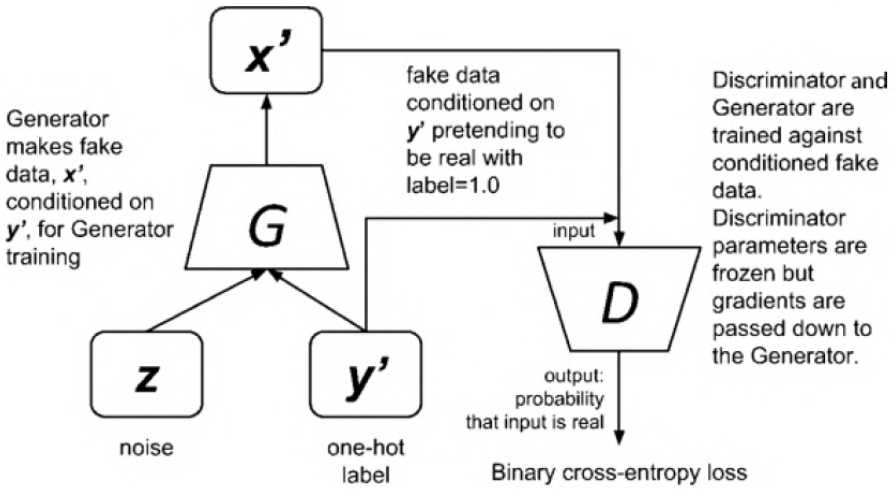


Figure 3.4: Training of the CGAN generator, where both the generated fake images are conditioned by a one-hot label [54]

The Auxiliary Classifier Generative Adversarial Network (ACGAN) is a type of CGAN, proposed by Odena et al., that differs from the original CGAN at the input and output levels [60]. The input to the discriminator is an image, whilst the output is the probability that the image is real and its class. Instead of feeding the side-information (label) to the network, which is done in

the CGAN, the ACGAN tries to reconstruct the side-information with an auxiliary class decoder network [54].

Besides the above mentioned GANs, there are other conditional GAN models, such as the InfoGAN [61] and the Semi-Supervised GAN [60].

3.1.3 Style GAN

The StyleGAN is a style-based GAN, meaning it is able to control several scale-specific features in an image, generating very high-quality images with control over very fine detail. This model can introduce higher levels of image sharpness and control than other conditional-based GANs.

The StyleGAN algorithm, firstly proposed by Karras et al. [62], is based on the original GAN architecture, with a resembling baseline configuration from the Progressive GAN (ProGAN) [63]. This algorithm has the fundamental aim of controlling the image synthesis process, motivated by style transfer, first proposed in [64].

Unlike most GAN architectures that attempt to improve the discriminator and its training, the StyleGAN aims at improving the generator network, even without modifying the discriminator or the loss function. The generator network starts with a learned input and adjusts the style at each convolution layer, based on the latent code, which allows it to control the weight of image features at a wide range of scales, modifying high-level attributes as well as stochastic features, such as hair strands and freckles on a human face [62].

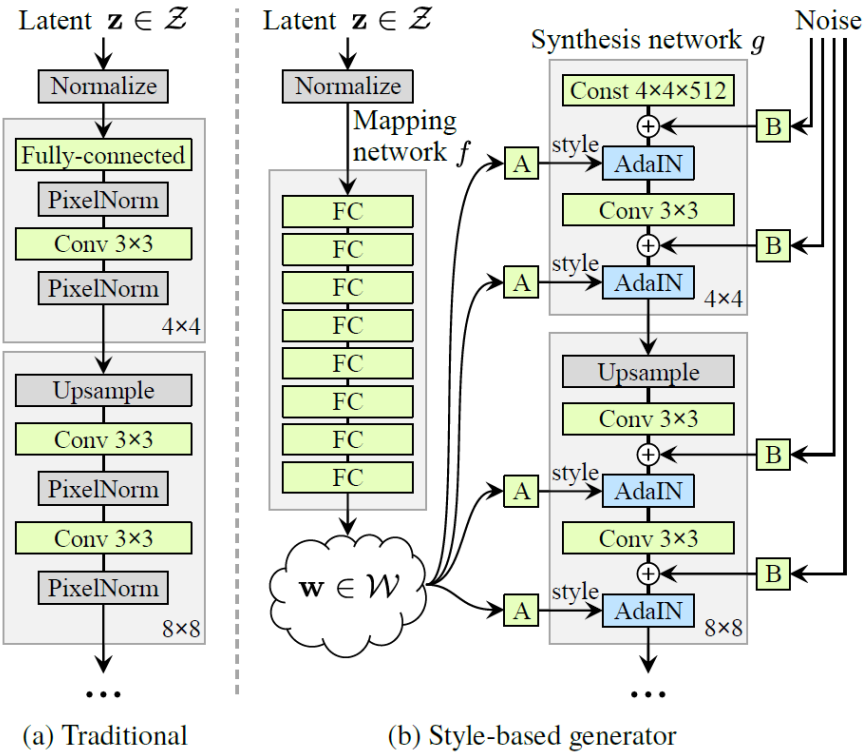
The common method of introducing the latent code in the generator is done through the input layer, however, in the StyleGAN, the input layer is omitted and replaced by a learned constant. The generator network embeds the input latent code, z, into an intermediate latent space, W, through a non-linear mapping network, which has a significant weight on how the factors of variation are represented in the network. The intermediate latent space W controls the generator through adaptive instance normalization (AdaIN) at each convolutional layer [65]. Figure 3.5 shows the architectural structure of the StyleGAN algorithm.

Each feature map x_i is separately normalized, scaled and biased using the corresponding scalar components from style y. AdaIN is defined by the following equation:

$$AdaIN\left(x_i, y\right) = y_{s,i} \frac{x_i - \mu\left(x_i\right)}{\sigma\left(x_i\right)} \tag{3.8}$$

where, x_i is the feature map, y is the style input, $\sigma(x_i)$ corresponds to the variance of the feature map input and $\mu(x_i)$ to the mean of the feature map input.

The input latent space must follow the probability density of the training data, which leads to a limited ability to control visual features with the input vector. This results in a unavoidable entanglement, due to the model not being capable of mapping parts of the input. However, the intermediate latent space does not follow the same data distribution and can reduce the correlation between features, therefore being allowed to be disentangled.



(a) Traditional (b) Style-based generator

Figure 3.5: Comparison between the architecture of a conventional GAN and a StyleGAN [66]

This algorithm also allows for mixing regularization where a two latent codes are used to generate a given percentage of images, instead of having just one latent code, during training. The process of switching from a latent code to another is referred to as *style mixing*.

As for the stochastic features, common GANs receive only one input of random noise through the input layer. Thus, the generators still have to find a way to generate spatially-varying pseudo-random numbers from earlier activations whenever they intend to implement stochastic variation. This additional process can cause a loss of control over other features due to feature entanglement. StyleGANs controls the stochastic variation by adding per-pixel noise after each convolution, which allow for available noise in each layer, disregarding the need for generating stochastic features from earlier activations, leading to a localized effect.

3.1.4 Few-shot Image Generation

State of the art GANs require large computational capacities, large training times and large datasets to perform at the desired level. Due to the demanding requirements by these large GAN variations, more compact and lighter GANs are sought after by the scientific community. Several GAN variations have been developed with the goal of allowing for faster and easier training with few training samples.

Some methods try to fine-tune larger or more complex models with images from the target domain and this task often shows better performance than a model trained from scratch with images

from the target domain. This is due to the fact that pre-trained models acquire useful weights that cannot be obtained using a small target dataset.

Nevertheless, due to the limitations of fine-tuning GANs, regarding the number of training images and distant source and target domains, the training of these models can often lead to poor results when on small datasets.

Lightweight GAN

The Lightweight GAN (LWGAN), originally developed by Liu et al., is a compact and light unconditional GAN that has the ability to converge from scratch with few training hours and small training sets. In the original paper, it was shown to be able to consistently gain superior quality on 1024×1024 resolution images [67], even when compared with state of the art GANs. The LWGAN is ideal for scenarios where the available datasets are small or with class imbalances and when the available computational power limits the array of trainable models. Additionally, fine-tuning pre-trained models can lead to worse performances due to bias, hence the need for a lightweight and compact GAN.

The demanding conditions of a small dataset and small computational power can lead to a high risk of over-fitting and mode-collapse. In order to avoid these setbacks, the generator has to learn fast and the discriminator has to continuously provide useful information to train the generator. The LWGAN variation differs from the original GAN in that it integrates skip-layer channel-wise excitation modules and a self-supervised discriminator as a feature-encoder.

Few-Shot GAN

The Few-Shot GAN (FSGAN) is a simple and effective method for adapting GANs in few-shot conditions (less than 100 images). FSGAN learns to modify the singular values of the pre-trained weights while freezing the corresponding singular vectors via repurposing component analysis techniques. This creates a large parameter space for adaptation while keeping changes to the pre-trained weights limited [68].

The FSGAN trains using a minimal number of training images from a new target domain, a method for adapting a pre-trained GAN to create unique, high-quality sample images. To do so, the number of trainable parameters is limited to a few highly expressive parameters that modulate orthogonal characteristics of the pre-trained weight space.

The network weights of a pretrained GAN (generator + discriminator) are first decomposed using singular value decomposition (SVD). With fixed left/right singular vectors, the singular values are then adapted using GAN optimization on the target few-shot domain. The authors demonstrate that changing single values in the weight space causes semantically relevant changes in the synthesized image while maintaining natural structure. The FSGAN has better picture quality after adaptation than other approaches that finetune all GAN weights [69], particular layers [70], or just alter batch norm statistics [71].

SinGAN

SinGAN is an unconditional generative model that can be learned from a single natural image [72]. It is trained to capture the internal distribution of patches within an image, and it can subsequently provide high-quality, diverse samples with the same visual content as the image.

The SinGAN is made up of a pyramid of fully convolutional GANs, each of which is in charge of learning the patch distribution at a different image scale. This enables the creation of new samples of any size and aspect ratio with high variability while maintaining the training image's global structure and fine textures. It allows to work with complex structures and textures without having to rely on a database of images belonging to the same class.

SSGAN

The SSGAN is a new strategy for moving the pre-trained generator's prior knowledge from a large dataset to a small dataset in a different domain [71]. The model can generate images based on this previous knowledge, which is something that can't be learned from a short sample. The strategy that focuses on the hidden layers in the generator's batch statistics, scale, and shift parameters. The GAN achieves stable training of the generator by simply training these parameters in a supervised manner, and the end model can generate higher quality photos than earlier approaches without collapsing, even when the dataset is smaller, with approximately 100 images.

The SSGAN focuses on the scale and shift parameters of batch statistics in the generator to adjust prior knowledge. The active filter in the convolution layer is controlled by these parameters, and adjusting the scale and shift parameters selects filters that are useful for generating images that are similar to the target domain. As a result, the SSGAN is a new generator transfer method that just updates the generator's scale and shift parameters. The number of images required to train the generator is decreased by updating only these parameters and fixing all kernel parameters in the generator.

3.1.5 Image-to-image Translation

While the previously mentioned GANs aim to generate images from a random noise input, image-to-image translation GANs aim at the conversion of an image of a certain domain into an image of a different domain [73].

Pix2Pix

The Pix2Pix algorithms bear resemblance to the previously mentioned CGAN, since in this case the condition of the image is to be translated. The working principle revolves around the idea of learning the mapping between an input and an output image using a set of training image pairs [74].

Pix2Pix is a generative model that applies a classification strategy to $N \times N$ patches of the input images. This model uses a U-Net [75] for the generator network and a Convolutional PatchGAN [76] for the discriminator network. Equation 3.9 shows the loss function of the Pix2Pix GAN and

Figure 3.6 shows an example of the Pix2Pix model. This algorithm is trained by optimizing the CGAN loss function with an additional factor to measure the similarity between corresponding real and generated images, L_1, which results in a minimized blurring in the generated image. The weight of the L_1 distance term is controlled by the hyper-parameter λ. As for the generator network, it was trained to maximize $log(D(x, G(x,z)))$, to prevent vanishing gradients.

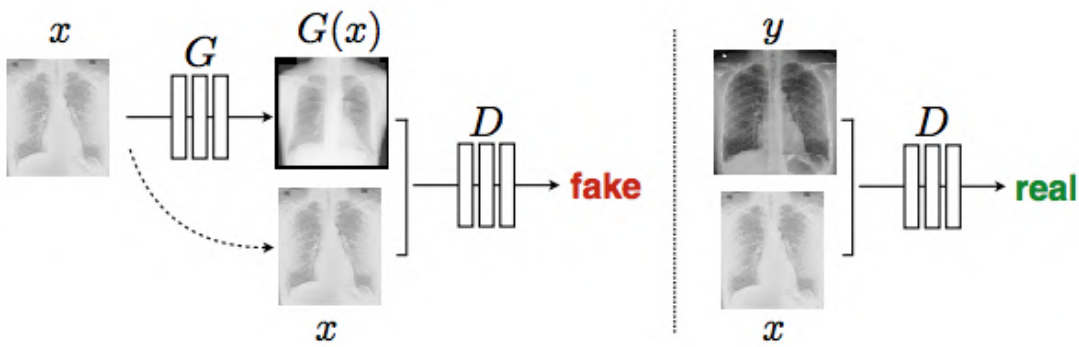


Figure 3.6: Example of Pix2Pix model translating an image from one domain to another

$$arg\min_{G}\max_{D}\mathbb{E}_{x,y}[logD(x,y)] + \mathbb{E}_{x,y}[log(1 - D(x, G(x,z))) + \lambda\mathbb{E}_{x,y,z}[||y - G(x,y)||_1] \tag{3.9}$$

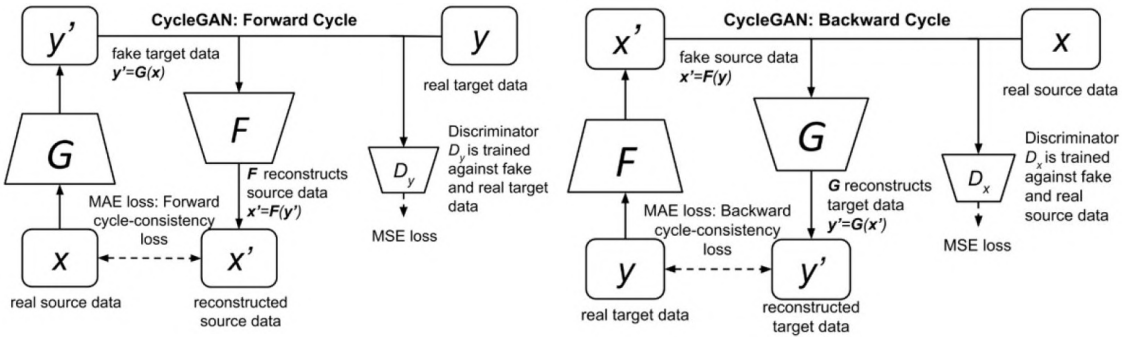
Noise is fed trough dropouts in the network, due to the generator being able to learn to ignore the noise when provided as a direct input. Additionally, the discriminator network's objective function is optimized to slow down the learning rate.

The major disadvantage of the Pix2Pix algorithm is the requirement of paired images, which can be expensive or even impossible to generate.

CycleGAN

Oppositely to the previous method, CycleGANs do not need paired images, for this unsupervised model can use uncorrelated images, as long as there is a sufficient amount and variation between source and target data [54].

This generative model is characterized by having two cycles, as shown in Figure 3.7, translated into four networks, of which two are generator networks, G and F, and the remaining two are discriminator networks, D_x and D_y.



(a) Forward Cycle of the CycleGAN

(b) Backward Cycle of the CycleGAN

Figure 3.7: Network model of the CycleGAN [54]

Generator G converts data from the source domain, x, to the target domain, y. This generator is trained by a GAN similar to the original GAN, presented in Figure 3.2a, where the discriminator network D_y is trained in the same adversarial structure. There is no need for supervised learning, since only real available images, x, are used in the source domain, and real images, y, in the target domain.

As shown in Figure 3.7, the forward cycle-consistency network guarantees that the real source data can be reconstructed from the fake data, which is done by minimizing the cycle-consistency L_1, defined by Equation 3.10. The cycle-consistency check assures that the source data x can be transformed to the domain y, maintaining the original features of x intact in y and being able to recover them. The backwards cycle of the CycleGAN is symmetric to the forward cycle, but the roles of the source data and target data are reversed. The source data becomes y and x the target data, as well as the generators G and F, in which F is just another generator network that takes part in the backward cycle. The backward cycle-consistency loss function is as shown in Equation 3.11. In the forward cycle, G was the generator network and F the network used to recover the data.

$$L_{forward-cyc} = \mathbb{E}_{x \sim p_{data}(x)}[||F(G(x)) - x||_1] \tag{3.10}$$

$$L_{backward-cyc} = \mathbb{E}_{y \sim p_{data}(y)}[||G(F(y)) - y||_1] \tag{3.11}$$

In summary, the ultimate goal of the CycleGAN is for the generator G to learn how to synthesize fake target data, y', fooling the discriminator, D_y, in the forward cycle. Since the network is symmetric, CycleGANs intend for the generator F to learn to synthesize fake source data, x', fooling the discriminator, D_x, in the backward cycle.

3.1.6 GAN Optimization

The training of GANs is still considered the most challenging aspect of GAN architecture, with it being an ongoing research topic. Diminished gradients, non-convergence and mode collapse are some of the common limitations with GAN training. In [77], several processes related to GAN

training have been highlighted as solutions to improve training and reduce the incidence of the mentioned limitations:

- Feature matching requires the generator network to generate data that matches the statistics of real data, making the generator match the expected value of the features on an intermediate level of the discriminator.
- Minibatch discrimination aims at avoiding mode collapse, by allowing the discriminator to look at multiple data examples in combination.
- Historical averaging tracks previous model parameters, penalizing changes that largely different from the average changes, which improves convergence.
- One-sided label smoothing replaces the binary classification with a smoothed classification, replacing 0 and 1 with values such as 0.1 and 0.9, which reduces the vulnerability of neural networks to adversarial examples [78]. Some authors advise not smoothing fake labels.
- Virtual batch normalization consists of normalizing each input sample to a referenced batch fixed statistic, which avoids the dependency on other inputs from the same minibatch.
- Progressive growing of GANs consists of progressively growing the generator and discriminator network, starting with lower resolution images and progressively improving training by adding layers to the networks for the higher resolution details. It starts by extracting the lower frequency information first and progressively decreasing the scale of the features and learning more fine scale information. The training process of each newly added layer is also progressive, with progression in the amount of images passed through the layer, until it is completely faded, as shown in Figure 3.8 [63].

Loss Functions

Since the original GAN appeared, dozens of newly developed GANs have been proposed. The majority of these GANs try to optimize the loss function by adding new penalties and other ways of computing the cost values. Besides the original formulation in Equation 3.1, the Wasserstein and Least Squares loss functions are the most commonly used for optimizing the learning phase in a large variety of GANs.

The Wasserstein GAN (wGAN) differs from the original GAN loss function in the discriminator network, as it does not classify instances into real or fake [79]. In the wGAN, the discriminator outputs a factor related to the "realness" of images, where a real images gets a higher value than a fake image. From this point on, the discriminator classifies the level of "realness" and not just a discriminating factor of either real or fake. The loss function of the wGAN results in the following equations for the discriminator network and generator network, respectively:

$$L^{(D)} = -\mathbb{E}_{x\sim p_{data}} D_w(x) + \mathbb{E}_z D_w(G(z)) \tag{3.12}$$

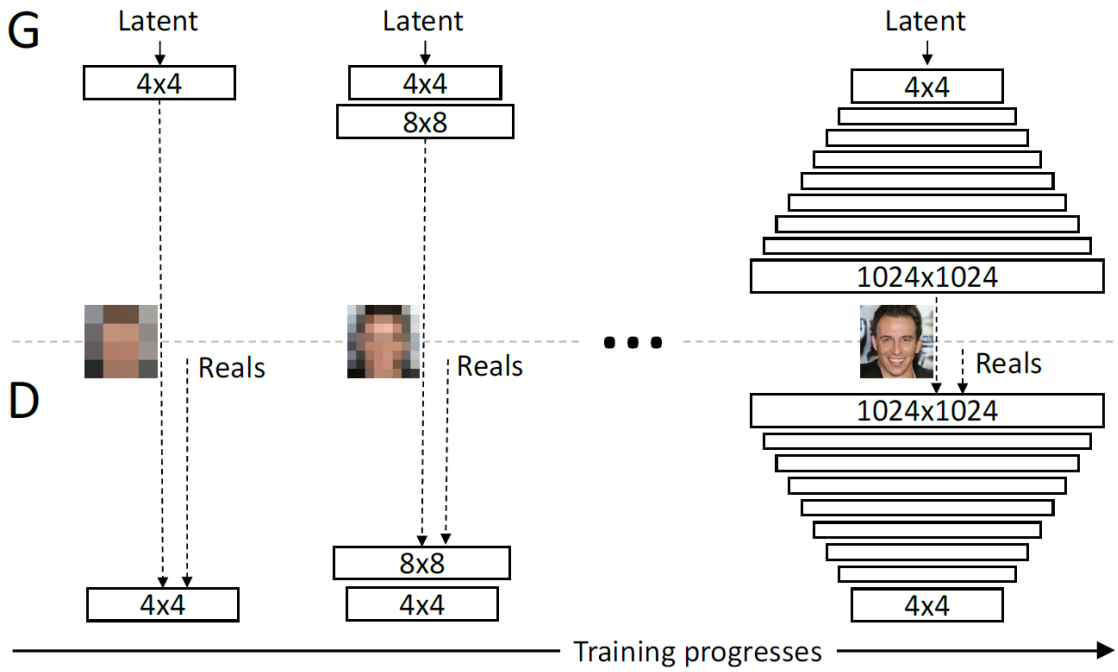


Figure 3.8: Progressive growing of both the generator and the discriminator network. Starting with lower resolution images and progressively improving training by adding layers to the networks for the higher resolution details. It extracts the lower frequency information first and progressively decreases the scale of the features, learning more fine scale information [63]

$$L^{(G)} = -\mathbb{E}_z D_w(G(z)) \tag{3.13}$$

The Wasserstein loss provides a useful gradient, allowing for a continued training of the models. A lower Wasserstein loss equates to a better fake image quality, which means that the generator network's goal is to minimize this function. Additionally, to prevent a vanishing gradient, the generator penalizes the generated images that fall too far form the real images. The discriminator network is modified to minimize the sum squared difference between predicted and expected values for real and artificial images:

$$\min_D (D(x) - 1)^2 + (D(G(z)))^2 \tag{3.14}$$

The generator network aims at minimizing the sum squared difference between predicted and expected values, as though the artificial images were real:

$$\min_G (D(G(z)) - 1)^2 \tag{3.15}$$

The Least Squares GAN (LSGAN) provides a loss function that penalizes larger errors, which results in a larger correction rather than a vanishing gradient and no model update [80]. For the least squares loss function, as long as the fake sample distribution is far from the real sample distribution, the gradients do not vanish. Thus, the generator network will keep trying to improve

its estimate of real density distribution even if the fake samples are on the correct side of the decision boundary [54].

The mean squared error aims at minimizing the generator loss function and, consequently, fooling the discriminator to incorrectly classify the generated data as real. Minimizing the discriminator loss function implies that the mean squared error between real data classification and the true label 1.0 should be close to zero.

$$L^{(D)} = \mathbb{E}_{x \sim p_{data}} (D(x) - 1)^2 + \mathbb{E}_z D(G(z))^2 \tag{3.16}$$

$$L^{(G)} = \mathbb{E}_z (D(G(z) - 1)^2 \tag{3.17}$$

3.2 Evaluation

The evaluation of GANs is still an underdeveloped area and an open problem, being one of the major barriers regarding the development of these algorithms. In order to properly evaluate a GAN's performance, it is important to choose evaluation metrics that cover the disadvantages of each other. Despite there being several evaluation measures, there is no consensus on which are most adequate and best translate the strengths and weaknesses of GAN models as well as evaluating the generated image quality. The most commonly used quantitative evaluation methods, such as FID and IS, are flawed and are therefore often used in conjunction for more comprehensive and complementary quantitative evaluation methods.

As defined before, the objective function of both the generator and discriminator networks measures the ability of these networks to perform against each other. While this can be a good measure for the overall GAN architecture, it isn't adequate to measure the quality and similarity of the generated images when compared to the real images. Thus, there is the need for a set of qualitative and quantitative metric measures.

As defined in Borji et al., qualitative measures, such as having a person distinguish between a real and an artificially generated image, can be useful for the evaluation of the image from the large scale point of view [81]. However, such measures may favor models that concentrate on limited sections of the data. On the opposite end, quantitative measures, while being less subjective, can relate to other aspects and be more efficient for the overall modeling.

According to [81], a good performance measure should:

- favor models that generate high fidelity samples i.e, the ability to distinguish generated from real samples,
- favor models that provide diversity of samples, being sensitive to overfitting, mode collapse and mode drop,
- favor models with disentangled latent space as well as space continuity,
- be sensitive to image-level transformations and distortions,

- agree with human perception and ranking of the models, and
- have low sample and computational complexity.

With the above characteristics in mind, several GAN evaluating measures have been developed and studied.

Qualitative Measures

Perceptual studies consist of asking humans to classify images as real or fake and evaluate the results to create a metric to quantify the quality of the model. Because this method is based on a subjective evaluation, the results can vary between annotators, the setup or by the use of hand-picked samples.

Another qualitative evaluation metric is the Visual Turing Test (VTT), firstly proposed by Geman et al.. It consists of a set of binary questions proposed by the system, regarding a set of images, and answered by the operator to assess the system's capability of recognizing objects and identifying attributes and relationships in images [82].

In [83], the authors performed a VTT to two radiologists to evaluate the generated samples of CT images with lung nodules for diagnosis of lung cancer, with the goal of evaluating the performance of the DCGAN that was developed to generate the images. Both radiologists had a mean inter-observer agreement for malignant and benign cases of 58.56% and for real and generated cases of 44.91%, which indicates the algorithm used had the capability of generating high-quality lung nodules.

Quantitative Measures

The Inception Score (IS) [77] is a classification performance metric and one of the most commonly used metrics for evaluating GAN models. It was initially proposed as a way to overcome the downsides of some qualitative methods subject to human evaluation. This metric uses a pre-trained neural network (Inception net) which is trained on ImageNet [37] to obtain the conditional label distribution $p(y|x)$. The IS measures two parameters simultaneously:

- Image variation, namely if the generated image is coming from a unique label.
- Meaningfulness of the objects, i.e., if each generated image uniquely looks like a meaningful sample.

For a given image output, if both the above statements are true, then a high IS is assigned and if one or both statements are not fulfilled, the IS is low, resulting in a low performance. The mathematical definition of the IS is shown in Equation 3.18:

$$IS = e^{\mathbb{E}_x D_{KL}(p(y|x)||p(y))} \tag{3.18}$$

where x is the sampled image, D_{KL} is the KL-divergence from the distributions, $p(y|x)$is the conditional class distribution and $p(y)$ is the marginal class distribution.

Despite the advantages of the IS, this evaluation metric still has some limitations [81, 84]:

- The IS is limited by the pre-trained network, being dependent on the classes of the ImageNet training set. If the classes of the generated image differ from the ones in the training set, the IS score is low.
- It is sensitive to small variations on the pre-trained weights, meaning the same test set can have different scores. The randomness associated with the training of a network leads to different training procedures to produce different weights.
- There is no difference to the IS if the generator replicates the training images or generates new ones, assigning high scores in either case.
- The IS has no evaluating measure for intra-class variability, meaning the IS will not be affected in the case of the generator only generating one type of image.
- The IS is an asymmetric measure, which is affected by image resolution.

The Fréchet Inception Distance (FID) [85] is a statistical metric and currently one of the state-of-the-art measures for evaluating GANs. It is strongly affected by mode collapse and has a relatively low variance, when given sufficient samples. The FID is based on the Inception network which extracts features from an intermediate layer, with which the FID between two multivariate Gaussian distributions with mean μ and covariance Σ is calculated, for both real and generated images, r and g, respectively. It can be calculated by the following equation:

$$FID = ||\mu_r - \mu_g||^2 + Tr(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{\frac{1}{2}}) \tag{3.19}$$

where r is the real data distribution, g the generated data distribution, μ_r and μ_g the mean of real and generated data, respectively, and Σ_r and Σ_g the covariance of the real and generated data, respectively.

The lower bound of the FID scale is zero and it has no upper bound, where a lower FID indicates a higher similarity between a real image and its corresponding synthetic counterpart, equating to a high quality artificial image. The FID score is more robust to noise, in other words, in case of a single generated image per class, the IS is high, however, the FID score does not follow the same logic and returns the inferior results achieved, making the FID adequate to evaluate image diversity.

The Kernel Inception Distance (KID) is a metric similar to the FID [86]. Also based on the Inception network, it measures the dissimilarity between two probability distributions P_r and P_g using samples drawn independently from each distribution. *i.e.*, computes the squared Maximum Mean Discrepancy (MMD) between the feature representations of real and generated images [87]. A lower value of the KID indicates a larger visual similarity between real and generated images, with a minimum value of zero. Unlike the FID, the KID is unbiased, which provides increased

reliability, especially when there are fewer test images than the dimensionality of the Inception features [88]. The KID is calculated by the following equation:

$$MMD(\mathbb{P}_r, \mathbb{P}_g) = (\mathbb{E}_{x_r, x_r' \sim \mathbb{P}_r, x_g, x_g' \sim \mathbb{P}_g}[k(x_r, x_r^{'}) - 2k(x_r, x_g^{'}) + k(x_g, x_g^{'})])^{\frac{1}{2}} \tag{3.20}$$

The FID evaluation metric is a good indicator for a GANs performance, however, it is calculated by first computing the 2048-feature vector resulting from the pool3 layer from the Inception V3 network, which is trained on the ImageNet dataset. This dataset is composed of millions of labeled RGB images, however, the images contained in the dataset may not have a representation of the image domain being used for the training of the GAN subjected to evaluation, meaning the feature vector may not be meaningful for the specific purpose of the task at hand. Nevertheless, the FID is still the metric used as benchmark for comparison with other architectures and methodologies.

The IS is also a benchmark metric used for comparison with other related works, however, it is not as accurate as the FID or other more recent metrics, such as the KID. Additionally, the IS is indifferent to intra-class variability and re-generation of training samples or new original data.

The KID is adequate for quantitatively evaluating GANs, since it is an accurate and unbiased metric, unlike the FID [86]. It provides a more adequate value which correctly translates the GAN's performance.

Precision measures the quality of the generated images and recall the proportion of real distribution over generated distribution. The IS covers precision but not recall, while the FID measures both precision and recall. The main limitation is that it is impractical to apply these scores for largely variant image datasets and their use is limited to evaluations on synthetic data [89].

The Likeness Score (LS) is a distance-based separability index created to evaluate three main aspects of GAN generated images [90]:

- Creativity: non-duplication of the real images. Checks for GAN overfitting.
- Inheritance: generated images should keep the same style as the real images, without compromising the creativity aspect.
- Diversity: Artificial images should always be different from each other.

LS uses the Euclidean distance to measure difference or similarity between images. Additionally, this metric could provide some degree of explainability regarding the three main aspects previously mentioned. Results showed the LS metric can effectively measure the performance of GANs with some degree of explainability and in a less computationally expensive process.

The 1-Nearest Neighbour Classifier (1-NNC) measurement consists of assessing whether two distributions are identical in two-sample tests, showcasing the generated images to the nearest real images in the training set [91]. Given two sample sets, the leave-one-out (LOO) accuracy is calculated. This parameter is calculated using the Euclidean distance, which is sensitive to minor perceptual disturbances. If the two distributions match, the 1-NNC classifier yields around a 50%

LOO. If the accuracy is lower than 50%, it means the GAN overfits, where 0% means a complete overlap of the images and the sample of the other set [92]. This metric is considered an appropriate measure to evaluate GANs, for it has the same advantages of other metrics and outputs a score in the interval [0,1], similar to the accuracy in classification problems.

Furthermore, to quantify the disentanglement of the latent space, two metrics - *perceptual path length* and *linear separability* - are used. The perceptual length metric measures the difference of two consecutive images by interpolating with two random inputs, where significant changes imply multiple features have changed, suggesting these might be entangled. As for the linear separability, it measures how well the latent-space points can be separated into two distinct sets through a linear hyperplane, making each set correspond to a specific binary attribute of the image [93].

In addition to these, there are many other commonly used evaluation metrics such as the Structural Similarity Index (SSIM) [94], the Average Log-Likelihood [55] and the Mode Score (MS) [95].

For the best evaluation of a GANs performance, both qualitative and quantitative metrics should be used, for they complement each other and provide a more complete evaluation process.

3.3 Applications in Chest Radiography

Despite GANs being a fairly recent type of generative models, as mentioned, the medical domain has a lot to gain from high quality generation of data due to the large barriers, such as the highly unbalanced datasets and the difficulty of obtaining medical data. With this in mind, over the last years, the number of developed papers regarding the application of GANs in chest radiography has significantly increased, as shown in Figure 3.9. Figure 3.10 shows the distribution of GAN related publications categorized according to the image modality.

A systematic review on the creation of artificial images for radiology applications using GANs, proposed in [96], presents the evolution of GANs in radiology.

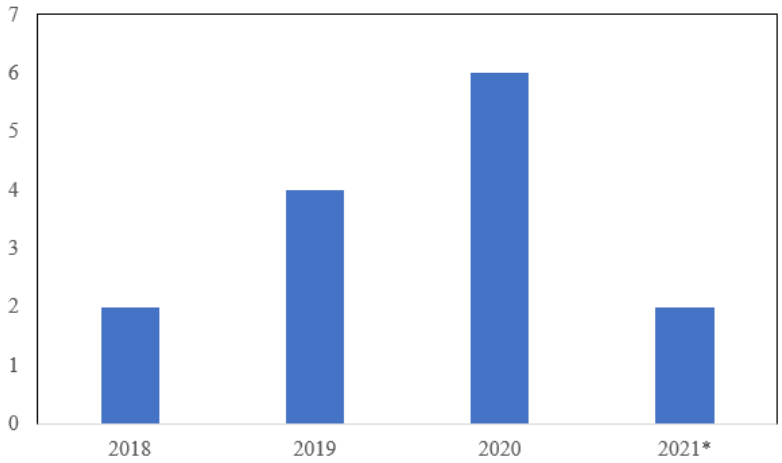


Figure 3.9: Evolution of GAN related publications in chest radiology between 2018 and 2021. Publications were searched for in the "Web of Science" website with the keywords: generative adversarial network; chest x-ray; GAN. Note that in 2021 only the publications published until February were taken into account [97]

Most of the GAN related publications available in medical imaging use either MRI or CT images. Since CXR is one of the most performed annual examinations, there is a disproportional representation of the CXR frequency in the available GAN publications. Still, GANs are useful for several types of applications in the medical domain such as data augmentation of datasets, translation of one image type to another and even improving the quality of existing images. In CXR applications, the majority of the developed GANs are used for data augmentation and only a small number is reported to be used for image translation.

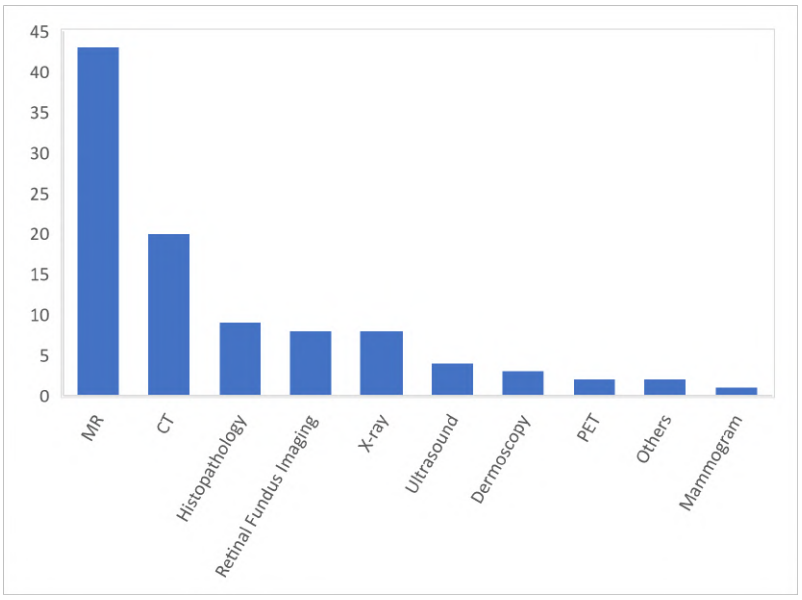


Figure 3.10: Categorization of GAN related papers according to image modality. The statistics presented are based on papers published on or before January 1st, 2019 [52]

Data Augmentation

In the work developed by Venu et al., a DCGAN is used for data augmentation of normal CXR images with the goal of training a deep neural network to classify the images as normal or as having pneumonia [98]. The dataset used in this work is highly unbalanced with a large predominance of the pneumonia class, which results in an over-fitted model, especially in the limited dataset that was used. The performance of the GAN was evaluated by the FID, resulting in a score of 1.289.

Waheed et al. propose a similar approach, for the detection of Covid-19 in CXR images [99]. Using an ACGAN based model, named CovidGAN, the authors aimed at generating CXR images to increase the available data to train deep learning algorithms for the classification and detection of pathologies related to Covid-19. The detection algorithm is based on CNNs, which can easily overfit in smaller datasets due to the large number of parameters. Classical data augmentation does not provide completely unseen images, therefore, synthetic data augmentation was used. Adam optimizer was used as the optimizer function during the training of the GAN, having the authors considered it the best choice for the optimization of the model. To evaluate the performance of the classification of both real and the generated images, the precision, recall, F1 score and sensitivity were calculated. However, no evaluation metrics were used to evaluate the GAN's performance. As for the classification algorithm, the accuracy achieved with the generated images was 95%, significantly improving the results.

The same logic was applied in the work developed by Salehinejad et al., where a DCGAN was used to artificially generate high resolution CXR images that mimic five common chest pathologies, to improve the distribution and size of an unbalanced dataset. A Deep Convolutional Neural Network (DCNN) was used to perform the chest pathology classification. For this purpose, the algorithms was based on the AlexNet, with some fundamental changes regarding the kernels, feature map sizes and convolutional layers. To optimize the learning phase of the DCGAN, an Adam optimizer was implemented. As for the evaluation, both quantitative and qualitative measures were used. The images were subjected to a qualitative analysis by a radiologist and the accuracy of each generated image was calculated, showing a significant improvement in the classification performance, when the generated images were used.

In [100] a DCGAN based generative model was used to generate artificial images of normal and abnormal cases containing cardiovascular abnormalities. The goal was to train an abnormality detector for cardiovascular pathologies, where a VGG-like structure was trained and used to classify the images. The classifier's performance was measured through the accuracy, which again proved to be improved with the presence of the generated images in the training set.

Image-to-image Translation

In the work developed by Liang et al., the authors propose two GANs for image-to-image translation between conventional and bone suppressed radiographs by Dual-Energy (DE) technique for chest radiographs [101]. The aim was to learn the mapping between conventional radiographs and bone suppressed radiographs, and classify thoracic pathologies. The two GANs are, respectively,

a Pix2Pix trained with patient-wisely paired radiographs, and a Cycle-GAN trained with unpaired radiographs. The purpose of using both GANs is to compare the effectiveness of using variations of GANs to suppress bone from standard CXRs. DE subtraction imaging captures several radiographs with different level energy, which are then combined to generate a bone suppressed image. Image-to-image translation is used to translate the standard radiograph into a soft tissue only radiograph [101]. The classification of the abnormality classification was done through the use of a VGG-19 [102] network. The chosen metric evaluation for this work consists of the Structural Similarity Index and Peak Signal-to-Noise Ratio (PSNR). In conclusion, the CycleGAN showed superior results over the Pix2Pix, which, according to the authors, is explained by this model not being restricted by paired CXRs, leading to a better generalization to unseen radiographs.

DeGrave et al. developed a CycleGAN to improve the explainability in the model's decision regarding the classification of Covid-19 [103]. While several very recent studies have been published regarding the high performance models applied to this subject, most of them lack the capability of providing an explanation for the model's decision. The commonly used machine learning classification methods capture features related to dataset-level differences, such as positioning of the patient and other markers, which can lead to undesirable shortcuts for the model's performance. Therefore, the work developed in this paper consisted of developing a GAN to transform Covid-19 negative samples into Covid-19 positive samples and vice-versa, in order to better comprehend what features could be used by machine learning algorithms to differentiate between the two types of samples. The rationale behind this was the fact that GANs are better at extracting all possible features that differentiate different datasets, instead of using saliency maps. As a result, the machine learning models used for the classifications were capable of predicting the GAN generated images as the images being transformed, demonstrating that the majority of features used by the classifiers were altered by the CycleGAN. The authors therefore concluded that machine learning models owe the majority of their models to the learning of shortcuts.

As described in this chapter, GANs show promising results in task of synthetic data augmentation and, with further development of optimization and training techniques, can achieve the desired purpose of generating high-quality artificial images.

Chapter 4

Methodology

4.1 Dataset

The data required for training generative models must be high-quality data, due to the model's high dependency on the training data for optimal performance. Some of the available datasets gather data from several sources and lack correct labels. With this in mind, the data used for training generative models in this project was collected from the VinDr-CXR dataset, due to its high-quality content.

4.1.1 VinDr-CXR Dataset

The VinDr-CXR dataset is a public dataset that was built with the intention of providing a large volume of CXR images with high-quality labels. The dataset consists of more than 100000 images in DICOM format. These images were collected from two major hospitals in Vietnam, Hospital 108 and Hanoi Medical University Hospital. Although the full dataset consists of more than 100 thousand images, the publicly available dataset is a subset of the full dataset, which contains 18000 postero-anterior (PA) view of CXR scans. The labels were manually annotated by at least three radiologists from a group of 17 radiologists with both the localization of critical findings and the classification of some common thoracic diseases [104].

The dataset is divided into 15000 images for the training set and 3000 images for the test set. The training set images were classified by three radiologists each, while the test set images were labeled by five radiologists. At the time of the development of this project, the test set labels were unavailable due to an ongoing competition that required the labels to be kept undisclosed.

The images of the VinDr-CXR dataset are divided into 22 critical findings, which are locally labeled with bounding boxes, and 6 global labels, with some examples shown in Figure 4.1. The resolution and format of the images is not fixed, varying in width and height, with a median value of 2788 by 2446 pixels.

	Characteristics	**Training Set**	**Test Set**
Collection Statistics	Years	2018 to 2020	2018 to 2020
	Number of scans	15 000	3 000
	Number of human annotators per scan	3	5
	Image size ($pixel \times pixel$, median)	2788×2446	2788×2394
	Age (years, median)*	43.77	31.80
	Male (%)*	52.21	55.90
	Female (%)*	47.79	44.10
	Data size (GB)	161	31.3
Local Labels	Aortic enlargement (%)	2348 (15.65%)	-
	Atelectasis (%)	62 (0.41%)	-
	Cardiomegaly (%)	1817 (12.11%)	-
	Calcification (%)	177 (1.18%)	-
	Clavicle fracture (%)	1 (0.01%)	-
	Consolidation (%)	121 (0.81%)	-
	Edema (%)	1 (0.01%)	-
	Emphysema (%)	14 (0.09%)	-
	Enlarged PA (%)	21 (0.14%)	-
	Interstitial lung disease (ILD) (%)	152 (1.01%)	-
	Infiltration (%)	245 (1.63%)	-
	Lung cavity (%)	21 (0.14%)	-
	Lung cyst (%)	4 (0.03%)	-
	Lung opacity (%)	547 (3.65%)	-
	Mediastinal shift (%)	85 (0.57%)	-
	Nodule/Mass (%)	410 (2.73%)	-
	Pulmonary fibrosis (%)	1017 (6.78%)	-
	Pneumothorax (%)	58 (0.39%)	-
	Pleural thickening(%)	882 (5.88%)	-
	Pleural effusion(%)	634 (4.23%)	-
	Rib fracture (%)	41 (0.27%)	-
	Other lesion (%)	363 (2.42%)	-
Global Labels	Lung tumor (%)	132 (0.88%)	-
	Pneumonia (%)	469 (3.13%)	-
	Tuberculosis (%)	479 (3.19%)	-
	Other diseases (%)	4002 (26.68%)	-
	COPD (%)	7 (0.05%)	-
	No finding (%)	10606 (70.71%)	-

Table 4.1: VinDr-CXR dataset and label prevalence.

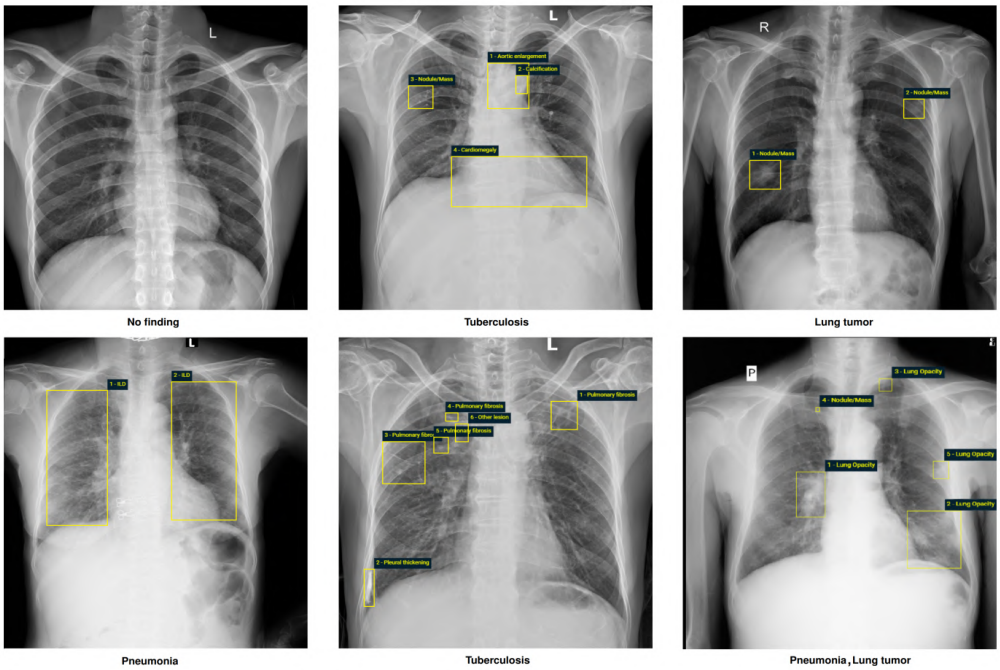


Figure 4.1: Examples of images found in the VinDr-CXR dataset with local labels by the bounding boxes and global labels listed at the bottom of each image [104]

4.2 Chest X-ray Image Generation

4.2.1 Lightweight GAN

State of the art GANs require large computational capacities, large training times and large datasets to perform at the desired level. Due to the demanding requirements by these large GAN variations, several GAN variations have been developed with the goal of allowing for faster and easier training. For this project, it was necessary to find a GAN that could use the available datasets and that could be trained with the few images available for some of the pathological classes. Furthermore, it was sought that this GAN was light and compact enough not to require large computational capacity, such as computing power and memory, allowing versatile training with the possibility of running several experimental tests in short periods of time. For this reason, the LWGAN was chosen as the GAN architecture used in this thesis.

The LWGAN architecture was based on the original publication [67]. The architecture consists of the original features, the SLE module and the self-supervised discriminator. The design was meant to be minimalist, using a single convolutional layer on each resolution in the generator and applying only three channels for the convolutional layers on resolutions of 512×512 in both the generator and the discriminator.

The LWGAN implementation was done with PyTorch [105] and all modifications were done in Python and through the use of other packages.

Figure 4.2 shows the structure of the skip-layer excitation module and the generator network of the LWGAN. The skip-layer excitation (SLE) module leverages low-scale activations to revise

the channel responses on high-scale feature-maps. This allows for a more robust gradient flow throughout the model weights for faster training. For the generation of higher resolution images, the generator is required to become deeper, which translates to more convolutional layers in concert with up-sampling needs. A deeper model leads to a larger number of model parameters and a weaker gradient flow through the generator, requiring larger training times [63, 106, 107].

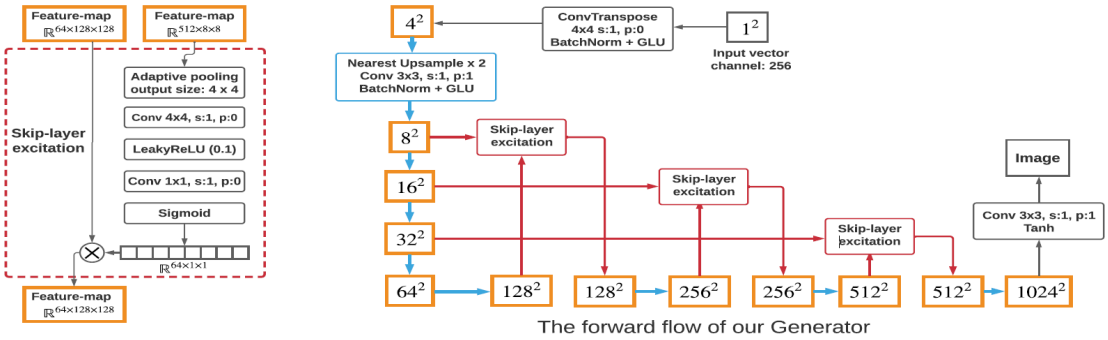


Figure 4.2: Generator network and skip-layer excitation module structures. Feature-maps are represented by the yellow boxes, the up-sampling structures are represented by blue boxes and arrows and the red boxes contain the SLE modules, as shown on the left [67]

Evidence shown in [102, 108] shows that network depth is of crucial importance, especially in computer vision tasks. A Residual Structure (ResBlock) is used to train the deep models by using a skip-layer connection as an element-wise addition between the activations from different conv-layers, however, as this required the spatial dimensions of the activations to be the same, channel-wise multiplications are applied between the activations, eliminating the heavy computation of convolution. The skip-connection idea is reformulated into the SLE module [109], which inherits the advantages of the ResBlock with a shortcut gradient flow, without the additional heavy computational burden.

The SLE module combines both the ResBlock and the Squeeze-and-Excitation (SE) module, proposed by Hu et al., which leads the LWGAN to benefit from the channel-wise feature re-calibration just as the SE, while strengthening the whole model's gradient flow with ResBlock [110]. According to the authors of the LWGAN, the channel-wise multiplication in SLE also coincides with Instance Normalization [65, 111], which is widely used in style transfer. The authors also show that SLE, similarly to the StyleGAN [66], enables the generator to automatically disentangle the content and style attributes.

As mentioned, the discriminator network in the LWGAN is a self-supervised discriminator. It is treated as an encoder paired with two small decoders, with which it is trained. It is a simple architecture, as shown in Figure 4.3, however, it has shown to be efficient and to provide a strong regularization for the discriminator. The auto-encoding forces the discriminator network to extract image features that the decoders can give good reconstructions on a smaller scale. For this, the discriminator is paired with the two decoders and optimized together in a reconstruction loss,

which is only trained on real samples, given by equation 4.1.

$$L_{recon} = \mathbb{E}_{f \sim D_{encode}(x), x \sim I_{real}}[||G(f) - \tau(x)||] \tag{4.1}$$

In this equation, f is the intermediate feature-maps from the discriminator, the function G contains the processing on f and the decoder, and the function τ represents the processing on sample x from real images I_{real}.

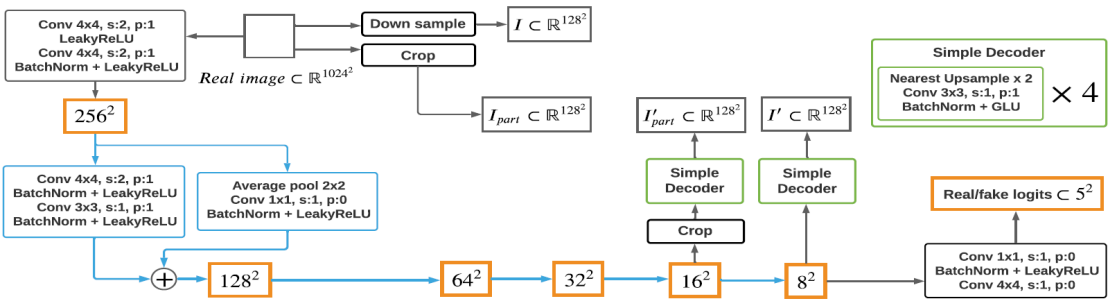


Figure 4.3: Structure and forward flow of the discriminator. Blue boxes represent the same residual down-sampling structure, green boxes represent the same decoder structure [67]

The two decoders are employed for two feature-maps with different scales: f1 on 16^2 and f2 on 8^2, and since the decoders only have four convolutional layers for the upsampling to 128×128, there are little extra computations involved. As described by the authors, f_1, the decoder on the feature-map 16×16, gets randomly cropped with $1/8$ of its height and width. The real image gets cropped on the same portion to get I_{part}, as shown in Figure 4.3. The real image also gets resized to get I. The decoders produce $I^{'}_{part}$ from the cropped f_1, and $I^{'}$ from f_2, the decoder on the feature-map of 8×8, from f_2. At the end the discriminator and the decoders are trained together to minimize the reconstruction loss in equation 4.1, by matching $I^{'}_{part}$ to I_{part} and $I^{'}$ to I.

The self-supervised strategy for the discriminator network allows it to be able to extract a more comprehensive representation of the inputs, since it is able to cover overall compositions, such as the features collected from f_2, and more detailed structures from the features collected by f_1. Other models and approaches aimed at improving a model's robustness and ability to generalize also employ similar methods involving auto-encoding approaches [112–115]. Additionally, the authors of the LWGAN also found that self-supervised discriminator network approaches significantly improve the synthesis quality of the generator network, among which auto-encoding showed the best performance boost.

With the addition of the SLE module and the two decoders, the LWGAN is able to effectively train with a small dataset and small batch-size, while assuring a fast training and avoiding over-fitting and mode-collapse.

The LWGAN implementation used in this thesis followed the original architecture. However, some of the parameters for the training of the model were altered from the start of the experimentation stage, such as changing the input images from RGB to greyscale.

4.2.2 Attention Modules

In addition, the implementation of the LWGAN used in this thesis is complemented by global self-attention. In neural networks, self-attention is a method that focuses on modeling long-range dependencies. Its superiority over other techniques, like convolution and recurrence, in terms of building global dependencies, has made it popular in modern deep learning [116]. Several recent efforts in computer vision have included global self-attention modules into Convolutional Neural Networks and showed promising results for various image understanding tasks [69, 117–120]. The high spatial dimensions of the input are the key problem when applying the global attention mechanism for computer vision applications. In a computer vision task, an input image often contains tens of thousands of pixels, and the attention mechanism's quadratic computational and memory complexity make global attention prohibitively expensive for such big inputs. The self-attention mechanism introduced in [116] is a global self-attention module, the GSA module, that performs attention while taking into account both the content and spatial placements of the pixels. The outputs of two parallel layers, a content attention layer and a positional attention layer, are summed at the end of the module. The content attention layer pays attention to all pixels at the same time, only on the basis of their content. It employs an efficient global attention method, comparable to [119, 120], with linear computational and memory complexity as the number of pixels increases. The attention map for each pixel is computed by the positional attention layer based on its own content and relative spatial positions to other pixels. As for the positional attention layer, a column-only attention layer precedes a row-only attention layer [116]. And so the used GSA module uses a non-axial global content attention mechanism that attends to the entire image at once rather than just a row or column. Figure 4.4 shows a representation of the GSA module.

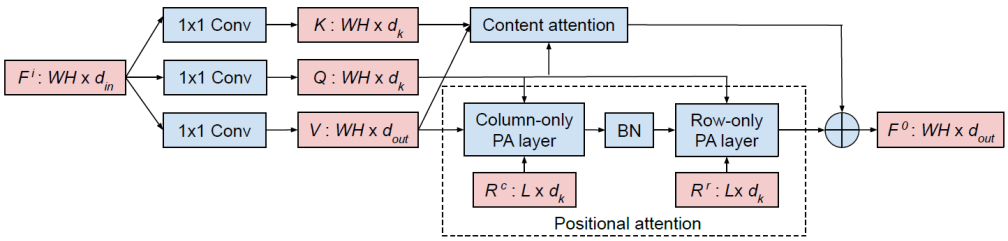


Figure 4.4: Representation of the GSA module [116]

The keys, queries, and values (constructed using 1x1 convolutions) are processed in parallel by the content attention and positional attention layers. The positional attention layer is divided into two sections: column-only and row-only, with learnt relative position embeddings R_c and R_r as keys. The output of the GSA module is generated by adding the outputs of the content and positional attention layers. Batch normalization [121] is denoted by *BN*, while positional attention is denoted by *PA*.

The GSA module is efficient enough to act as the backbone component of a deep network and so a global attention-based network composed of GSA modules instead of spatial convolutions to model pixel interactions is possible. In the implementation of the LWGAN in this thesis, the GSA

module and GSA network was available for use to try to further improve the training of the GAN model.

4.2.3 Loss Functions

The loss functions used by GANs varies with each architecture, however, for some architectures the importance of the loss function is more or less significant than for others. In the case of the original implementation of the LWGAN, a simple hinge version of the adversarial loss was used and the same loss was used for the majority of the experimental phase of this thesis [122, 123]. The hinge loss is defined by the following equations:

$$L_D = -\mathbb{E}_{x\sim I_{real}}[min(0,-1+D(x))] - \mathbb{E}_{\hat{x}\sim G(z)}[min(0,-1-D(\hat{x}))] + L_{recon} \tag{4.2}$$

$$L_G = -\mathbb{E}_{z\sim\mathcal{N}}[D(G(z))] \tag{4.3}$$

According to the authors of the LWGAN and other approaches, different loss functions do not necessarily have a large contribution for the performance of the GAN and the overall training. The reason for the hinge loss as the chosen loss function lies with the fact that it computes the fastest.

Another loss function used for this project is the dual contrastive loss function [124]. The authors of the dual contrastive loss associate the replacement of loss functions, such as the logistic loss from the StyleGAN2 implementation, with an attention mechanism. Models coupled to attention with their re-weighting mechanisms provide a possibility for long-range modeling across distant image regions. Contrastive learning associates data points and their positive examples and disassociates the other points within the dataset which are referred to as negative examples, i.e., targeting a transformation of inputs into an embedding where associated signals are brought together and distanced from the other samples in the dataset. This type of learning has been shown to be an effective tool for unsupervised learning and in generative models [125–127]. According to the authors, the contrastive loss function aims at combining the teaching of the discriminator network to disassociate a single real image against a batch of generated images with the learning to disassociate a single generated image against a batch of real images, as shown in Figure 4.5.

As for the generator network, it tries to minimize the dual contrasts. The equations for both cases in the dual contrastive loss function are the following:

$$L^{contr}_{real}(G,D) = \mathbb{E}_{x\sim p(x)}[log\frac{e^{D(x)}}{e^{D(x)} + \Sigma_{z\sim\mathcal{N}(0,I_d)}e^{D(G(z))}}] \tag{4.4}$$

$$L^{contr}_{fake}(G,D) = \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[log\frac{e^{-D(G(z))}}{e^{-D(G(z))} + \Sigma_{x\sim p(x)}e^{-D(x)}}] \tag{4.5}$$

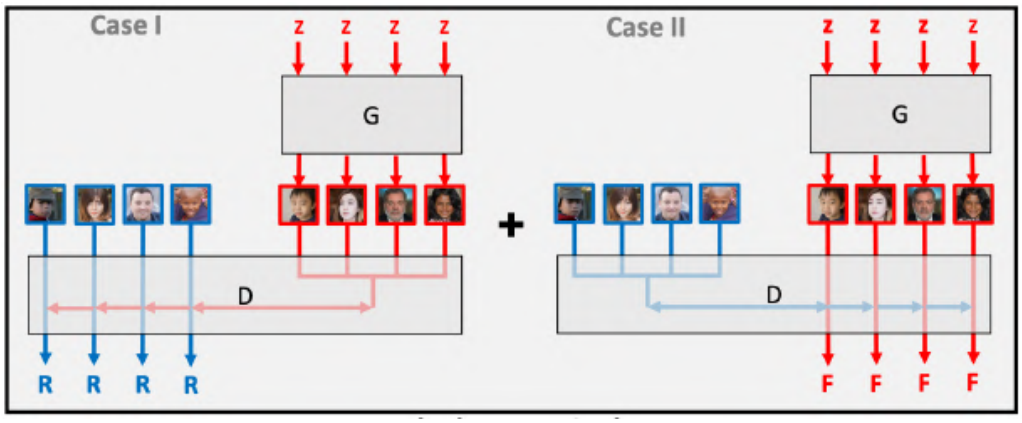


Figure 4.5: Architectural representation of both actuating parts of the dual contrastive loss in the discriminator network [124]

When compared with other loss functions, the dual contrastive loss function outperformed other losses on four out of five different datasets [124]. Additionally, the authors of the dual contrastive loss found the dual contrastive features to be consistently more distinguishable than the original discriminator features, which back-propagates more effective gradients to incentivize the generator network.

4.3 Quantitative Evaluation

On each evaluation checkpoint during training, the FID was calculated by comparing a set of 5000 real CXR images with a set of 5000 generated CXR images. Upon the final iteration of the training stage, the best model was selected, based on the FID values calculated throughout training, and the 5000 images were generated for the calculation of all three evaluation metrics, FID, KID and IS. The FID and KID were calculated using generated artificial CXR images and real CXR images from a set which did not contain any images used for training. The IS was calculated using only the set of generated artificial CXR images.

Besides evaluation metrics, other evaluation methods were used to assess the quality of the generated images. One of the methods was to run a set of artificially generated images in a binary classification model trained to classify CXR images as normal or abnormal. The goal was to evaluate whether the artificial images were considered normal in a more objective and computational evaluation, complementing the qualitative evaluation methods. The other quantitative evaluation method to be used consisted of comparing a classification network's performance when trained with a set of real images with the performance of the same classification network when trained with both real and artificial images. The goal was to assess if the artificial images had enough quality to be used for training in computer vision models.

4.4 Qualitative Evaluation

To complement the quantitative evaluation metrics, qualitative evaluation methods were also used, since these evaluation methods provide an insight on different aspects of the generated images, such as the human perception of the image quality and structure. To evaluate the images generated by the LWGAN, artificial images were submitted to be perceptually validated.

The perceptual validation consists in submitting generated images to be evaluated by radiologists. Comparison of real CXR images with artificially generated CXR images can provide a good indication of image quality at the overall image and structural level as well as at the detailed level of the images with regard to detail and resolution.

Figure 4.6: Platform for evaluation of CXR images

The goal was to assess the quality of the images and the overall similarity between real and artificially generated images from a human perspective and collect feedback regarding the decisions.

For this validation step, an evaluation platform was required so that participants could evaluate the images at random and anonymously, in order to avoid bias led by indicators regarding the source of the images.

This was done using an in-house software which presented a randomly selected subset of images. The platform allows for a window center/width adjustment, zooming and panning. Additionally, it is also possible to draw rectangles of any size on the image, covering any labels or marks, while saving the corresponding coordinates.

Figure 4.6 shows an example of a randomly selected image ready for classification by the user. The participant selects one option for each class, image source and pathology classification, and proceeds to the next image. Once all the images have been assessed, the results are automatically stored and ready for analysis.

Chapter 5

Experiments

This chapter presents the description and evolution flow of the development of the LWGAN applied to CXR images for artificial image generation. The approach taken and the goal for each test will be described and analysed in this chapter.

5.1 Data Preparation

For the training of the LWGAN, the images in the training set had to be prepared to be used in the model. 4000 images were randomly selected from the available images labeled as *No Finding* in the VinDr-CXR dataset. All of the performed tests used the complete set of the 4000 images or smaller subsets extracted from the training set. This value was chosen due to the fact that the pathological class with the largest number of images has 4000 samples, and there are a considerable amount of classes that have between 1000 and and 4000 images. This would allow for the replication of the training conditions if there was a need to train the LWGAN with pathological images.

5.2 Chest X-ray Image Generation

In this section, the experiments and ideology behind them will be explored along the four main areas.

This stage of the development of the LWGAN focused on experimenting with four main areas: resolution, self-attention, loss functions and hyper-parameters. Hyper-parameters, such as the training batch size, the augmentation technique types and augmentation probability were empirically adjusted according to initial experiments (cf. Hyper-parameters on Section 5.2.5). This adjustment was based on hardware limitations, such as GPU memory, and the results of tests carried out throughout the development of the LWGAN.

5.2.1 Resolution

Image resolution is an important factor when dealing with image classification models. Ideally, the resolution of training images is on par with the resolution of the real test images. However, increased image sizes lead to an increased computational cost and time, and decreases the maximum possible batch sizes for training. In this experiment, two image resolutions were tested: 256×256 and 512×512. Smaller resolutions, such as 128×128 were not relevant, since these may not provide sufficient detail for pathology diagnosis or description. Furthermore, and as shown by [128], the best results for image classification models in radiography were obtained with image resolutions larger or equal to 256×256 for the training images.

At this point, no attention mechanisms were added to the model, the hinge loss was used and the hyper-parameters were set as shown in Table 5.1.

Test	Image Resolution	Hyper-parameters and other specifications
256 Resolution	256×256	Batch-size: 6, 2000 images Augmentation types: Translation, Cutout, Color Augmentation Probability: 0.25
512 Resolution	512×512	Batch-size: 6, 2000 images Augmentation types: Translation, Cutout, Color Augmentation Probability: 0.25

Table 5.1: Image resolution comparison test

5.2.2 Self-Attention

From this point on, the resolution chosen for all of the experiments was 512×512, unless specified otherwise in the detailed description of each test, since besides obtaining the best results, it is not limited by the computational cost with regard to new additions and manipulations of the architecture. .

In order to further improve the training of the LWGAN, several tests were performed with the addition of attention in one or multiple layers of the model. These models were compared to the benchmark, which consists of the same architecture without the addition of attention in any layer.

The tests shown in Table 5.2 were performed with the goal of comparing different implementations of self-attention in the network's layers. The first test to be carried out trained a model with self-attention in the 32×32 layer with an image resolution of 512×512. The following tests consisted of experimenting with self-attention in different layers, such as 256×256. Later on, tests with attention in multiple layers were performed, such as the combination of the first three layers, 32×32, 64×64 and 128×128, and also a combination of self-attention in every layer up until the largest resolution.

The goal of carrying out a large amount of tests is to provide a solid foundation for the understanding of the influence of self-attention in the LWGAN architecture.

Test	GSA Layers	Hyper-parameters and other specifications
32 GSA	32×32	Batch-size: 6, 2000 images Augmentation types: Translation, Cutout, Color Augmentation Probability: 0.25
256 GSA	256×256	Batch-size: 1, 4000 images Augmentation types: Translation, Cutout, Color Augmentation Probability: 0.25
32-128 GSA	32×32 64×64 128×128	Batch-size: 6, 2000 images Augmentation types: Translation, Cutout, Color Augmentation Probability: 0.25
All Layers	32×32 64×64 128×128 256×256 512×512	Batch-size: 6, 4000 images Augmentation types: Translation, Cutout, Color Augmentation Probability: 0.25

Table 5.2: Experiments done with with the goal of evaluating the influence of self-attention in the training of the LWGAN

5.2.3 Loss Functions

In order to compare the performance of the hinge loss against the dual contrastive loss, two models were trained, one with each loss function. Both tests were performed with attention in one of the architecture's layers, namely the 32×32 layer. In the case of the hinge loss, the test described as *32 GSA* (Table 5.2) was the one used to compare against the dual contrastive loss test. Table 5.3 shows the model parameters and settings the test.

Test	GSA Layers	Hyper-parameters and other specifications
Dual Contrastive Loss	32×32	Batch-size: 6, 2000 images Augmentation types: Translation, Cutout, Color Augmentation Probability: 0.25

Table 5.3: Executed experiment for the comparison of a different loss function against the loss function used in the remaining tests

5.2.4 Large Resolution

A larger resolution model was trained to evaluate the GAN's capacity of training and generating high-resolution images. The images found in the VinDr-CXR dataset, and the ones commonly found in clinical practice, are high-resolution images, in this case with a median resolution of 2788×2446 pixels. Additionally, during the diagnosis process by radiologists, very small details and structures are used to identify and correctly diagnose pathologies. With this in mind, having larger resolution images being generated by GANs is ideal for applications such as pathology classification models. Nevertheless, larger resolution images require heavy computational capabilities and were, therefore, only used in this experiment.

To evaluate this test, it was compared against a similar architecture trained model that just differed in the resolution. Both tests were performed with attention in one of the architecture's

layers, namely the 32×32 layer. The test used for comparison was the *32 GSA* model (Table 5.2), as it already met the requirements of the test for the remaining parameters. Table 5.4 shows the model parameters and settings for the *large resolution* test.

Test	GSA Layers	Hyper-parameters and other specifications
1024 Resolution 1024×1024	32×32	Batch-size: 6, 2000 images Augmentation types: Translation, Cutout, Color Augmentation Probability: 0.25

Table 5.4: Experiment done with with the goal of evaluating the difference in training with large resolution images with the same hyper-parameters as the best performing model

5.2.5 Hyper-parameters

There are several parameters that underwent changes in addition to those previously mentioned. Most of these changes were made to hyper-parameters, as stated in this section.

Data augmentation techniques were used to increase the amount of training data provided for the model. These techniques include image translation, horizontal flip, contrast, brightness and saturation changes, image cutouts and offset, and are all randomly applied. Since CXR images are content-sensitive images, introducing random augmentations to the training set could negatively affect training. For example, setting a probability $p = 0.5$ for the horizontal flip augmentation, which is common practice in image augmentation strategies during training, in CXR images can lead to a wrong training of the model, since these images are sensitive to structure placement, i.e. the heart is almost always on the right side of the image (left side of the patient). With this in mind, only translation, horizontal flip, cutout and color related augmentation techniques were applied for all tests, taking into account the data augmentation probability. For all of the performed tests, the augmentation probability, ranging from 0 to 1, was set at 0.25 including for the horizontal flip augmentation.

The batch size of training images is dependent on the computational power and the computational cost of the model being trained. As a result, the batch size was limited for some of the larger models due to the hardware limitation. Nevertheless, to work around this limitation, the LWGAN is fitted with the option to update the gradient after a specific number of batches, meaning the gradient accumulates with values from past batches until it reaches a specific number of batches set as a parameter of the GAN. For this specific case, the batch size was kept at 6 images for most tests and the gradient accumulation feature at 12 batch sizes, totaling at 72 images. For the tests where the batch size could not be as large as 6, the gradient accumulation feature could be increased to achieve the same result.

Lastly, the number of random images available for training was also subject to experimentation. The original training set is composed of 4000 randomly selected images from the *No Finding* class of the VinDr-CXR dataset. Two additional training sets were created, one with 1000 randomly selected images from the 4000 image training set and another one with 2000 randomly

selected images from the same source. The goal was to assess the influence of the size of the training dataset on the performance of the LWGAN.

Collecting the above mentioned tests, Table 5.5 shows the detailed architecture and parameter selection for each test.

Test	GSA Layers	Hyper-parameters and other specifications
1000 Image Set	32×32	Batch-size: 6, 1000 images Augmentation types: Translation, Cutout, Color Augmentation Probability: 0.25
4000 Image Set	32×32	Batch-size: 6, 4000 images Augmentation types: Translation, Cutout, Color Augmentation Probability: 0.25
4000 Image Set Larger Batch Size	32×32	Batch-size: 32, 4000 images Augmentation types: Translation, Cutout, Color Augmentation Probability: 0.25
4000 Image Set Horizontal Flip $p = 0.5$	32×32	Batch-size: 6, 4000 images Augmentation types: Translation, Cutout, Color Augmentation Probability: 0.25

Table 5.5: Hyper-parameter variations in tests with similar architecture and 512×512 resolution

5.3 GAN Validation

5.3.1 Quantitative Metric Evaluation

Each test was run for a maximum amount of 150 epochs, however, all of the performed tests collapsed and diverged before they could reach that checkpoint. The FID was calculated every epoch for each test, which allowed for a continuous control of the performance and training. The results shown in Chapter 6 are the results for the best checkpoint of each model's training, which was then used to calculate the three evaluation metrics. For the calculation of the FID and KID, 5000 real images and 5000 artificially generated images were used. As for the IS, only the 5000 artificially generated images were used.

The artificial images were generated using a parameter averaging of the generator, Exponential Moving Average (EMA), which according to [129] improves the overall results, as shown also in this LWGAN's results. With this in mind, every image set for evaluation was artificially generated with this method.

5.3.2 Perceptual Validation

For the evaluation of the LWGAN, a mixed set of real images and artificially generated images was created and sent to six individuals: two radiologists, two PhD students in medical image analysis acquainted with CXR images and two individuals unacquainted and inexperienced with analysis of CXR images. This allowed to evaluate the level of knowledge required and the influence of the absence of knowledge regarding CXR images in distinguishing real from generated images. The

images were evaluated with regard to authenticity of the image and also, for the two radiologists, a binary classification of normal or pathological.

A total of 100 randomly selected images were submitted for validation, 50 artificial normal images, 25 real normal images and 25 real pathological images. The artificial images were generated by the *32 GSA* model.

5.3.3 Binary Classification Model

In order to establish if the generated images contained features representative of normal CXR images, an image set of generated normal CXRs was submitted for classification by a binary CXR classification algorithm.

The artificially generated image set was compared against the real image set from the VinDr-CXR dataset using the binary classification model. This model is based on a MobileNet [130] and was trained on three folds of, on average, 2121 normal images and 878 pathological images from the VinDr-CXR dataset. Two other sets were used for validation and testing. The validation set was used to calculate the classification threshold, 0.42, which is the best operating point of the RoC curve. The 6000 artificially generated normal images were submitted for classification and another 2120 normal real images were randomly selected for a comparison test. None of the real images submitted for classification were part of the training set for the classification model.

Each image was awarded a probability associated with the certainty of its predicted class. Images classified as normal got a probability close to zero and images classified as abnormal got a probability close to one.

5.3.4 Training of a Pathology Classifier

In order to establish if the generated images were of sufficient quality for the training of deep learning models, a YOLOv5 [131] object detection network was used. For this purpose, the 15000 CXRs of VinDR-CXR were randomly divided into train (60%), validation (20%) and test (20%) sets, preserving the approximate prevalence of each pathology as much as possible between the three divisions. Additionally, 6000 CXRs were artificially generated using the *32 GSA* model of the LWGAN, whose parameters are shown in Table 5.2. Three different training strategies were used:

- First, using only real pathological images;
- Second, using all real images, both normal and pathological;
- Third, using both real pathological images and artificially generated normal images.

Chapter 6

Results

6.1 Chest X-ray Image Generation

6.1.1 Resolution

Table 6.1 shows the results of the resolution related experimental LWGAN tests with regard to the IS, FID and KID.

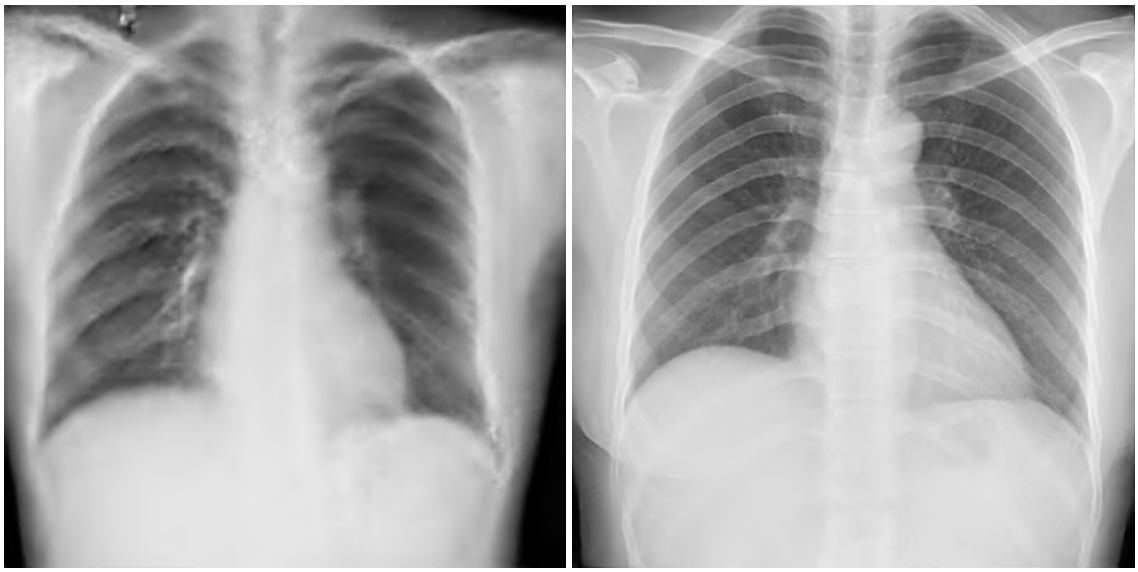
Test	IS	FID	KID
256 Resolution	1.895 ± 0.038	123.18	0.16380 ± 0.00354
512 Resolution	$\mathbf{2.047 \pm 0.0267}$	**24.13**	$\mathbf{0.01743 \pm 0.00083}$

Table 6.1: Resolution-related experimental tests and respective quantitative metric results for artificial CXR image generation. Bold values in each column indicate the best result for each metric.

Comparing the initial tests with different resolutions, the LWGAN trained with larger resolution images, 512×512, achieved a significantly better performance in all three evaluation metrics, shown in Table 6.1. Figure 6.1 shows two artificially generated samples from the two LWGAN models in Table 6.1, trained with different resolution images, 256×256 and 512×512. The images shown support the corresponding metric results for each image regarding its quality.

6.1.2 GSA

Regarding the experiments related to self-attention in the layers, shown in Table 6.2, the overall best result was achieved with the *32 GSA* test, in spite of the IS value for the *256 GSA* test being superior.



(a) 256×256 (b) 512×512

Figure 6.1: Artificially generated CXRs with different image resolutions

Test	IS	FID	KID
32 GSA	2.109 ± 0.034	**17.83**	$\mathbf{0.01211 \pm 0.00072}$
256 GSA	2.317 ± 0.040	65.61	0.06771 ± 0.00175
32-128 GSA	$\mathbf{2.318 \pm 0.051}$	77.22	0.08154 ± 0.00195
All Layers	2.065 ± 0.047	52.39	0.05846 ± 0.00169

Table 6.2: GSA-related experimental tests and respective quantitative metric results throughout the development of the LWGAN for artificial CXR image generation. Bold values in each column indicate the best result for each metric.

Figure 6.2 show two artificially generated samples by the *32 GSA*. Both images show a very clear and fine detail of the structures as well as an overall correct representation of the anatomical structures of a CXR.

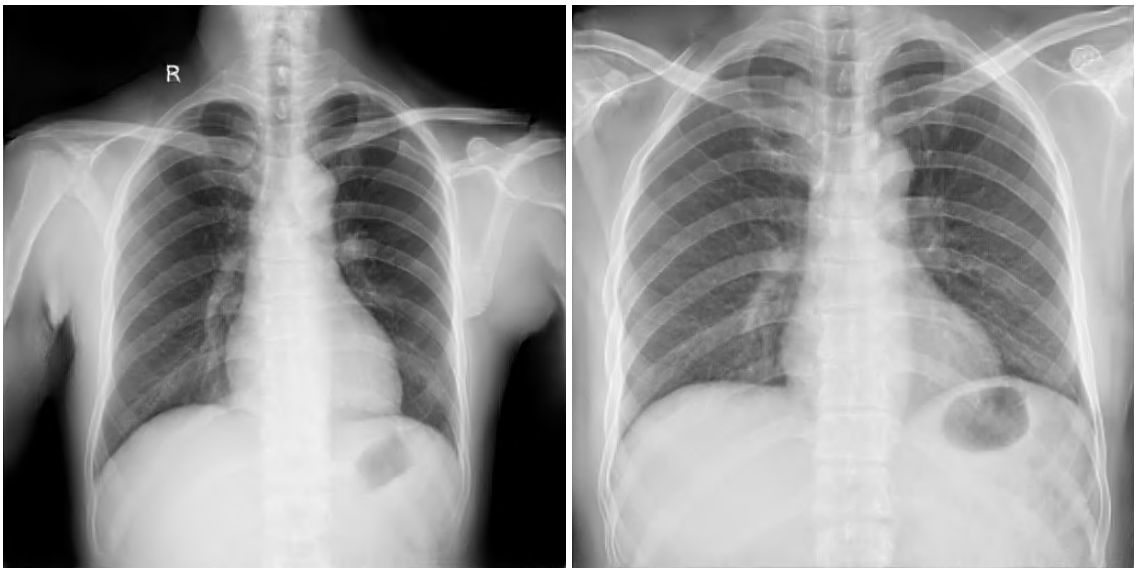


Figure 6.2: Artificially generated CXR samples from the *32 GSA* model

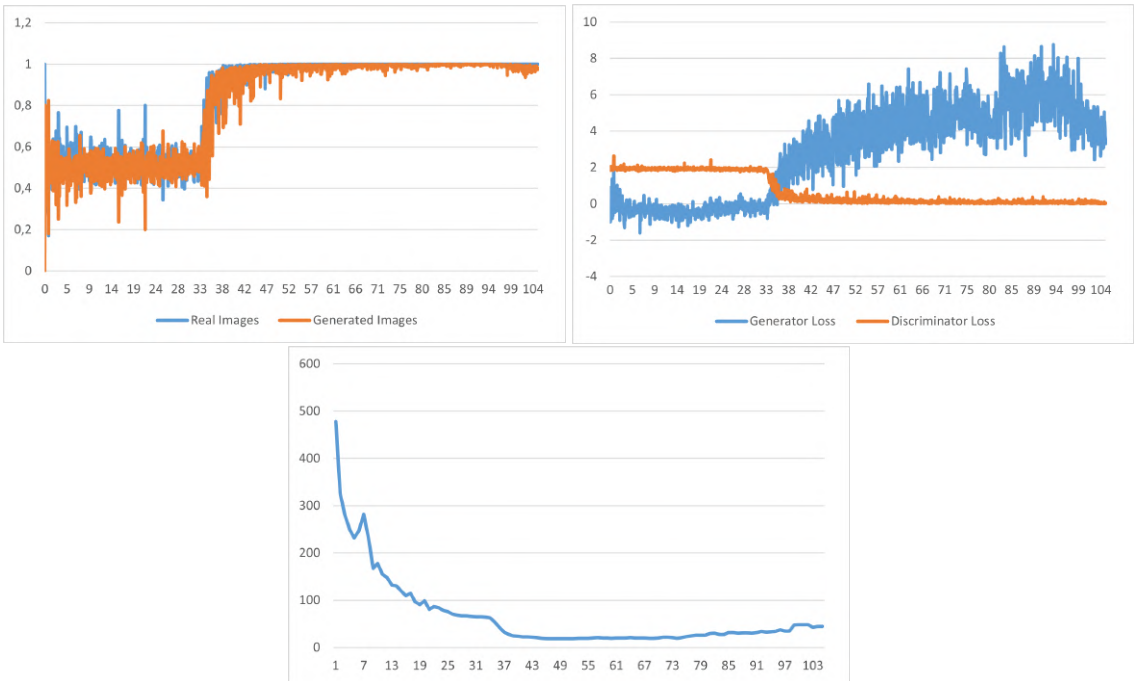


Figure 6.3: Discriminator accuracy at predicting the image source (left), loss functions (right) and FID (center) values from the *32 GSA* model throughout training

Figure 6.3 shows the evolution of the discriminator's accuracy, the generator and discriminator loss functions and the FID evolution, from the previously mentioned trained model throughout the development. Looking at the accuracy from the discriminator's predictions (Figure 6.3), the *32 GSA's* discriminator network was able to correctly predict the origin of each image with certainty. Additionally, all three graphs show a behaviour change at around epoch 33.

Figure 6.4 shows two samples generated by the *256 GSA* model. By including GSA in the

256×256 layer, the network learned to generate images with very different overall pixel intensities, which led to poor quality images.

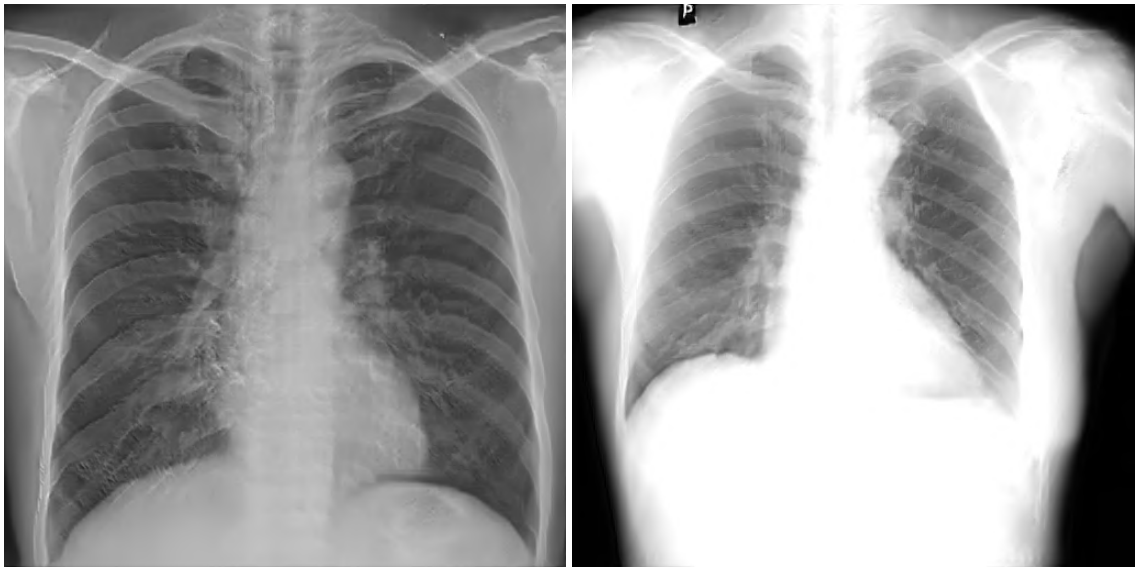


Figure 6.4: Artificially generated CXRs from the *256 GSA* model

Additionally, the overall detail is not sharp and the anatomical structures have some imperfections such as ripples in the edges of the ribs.

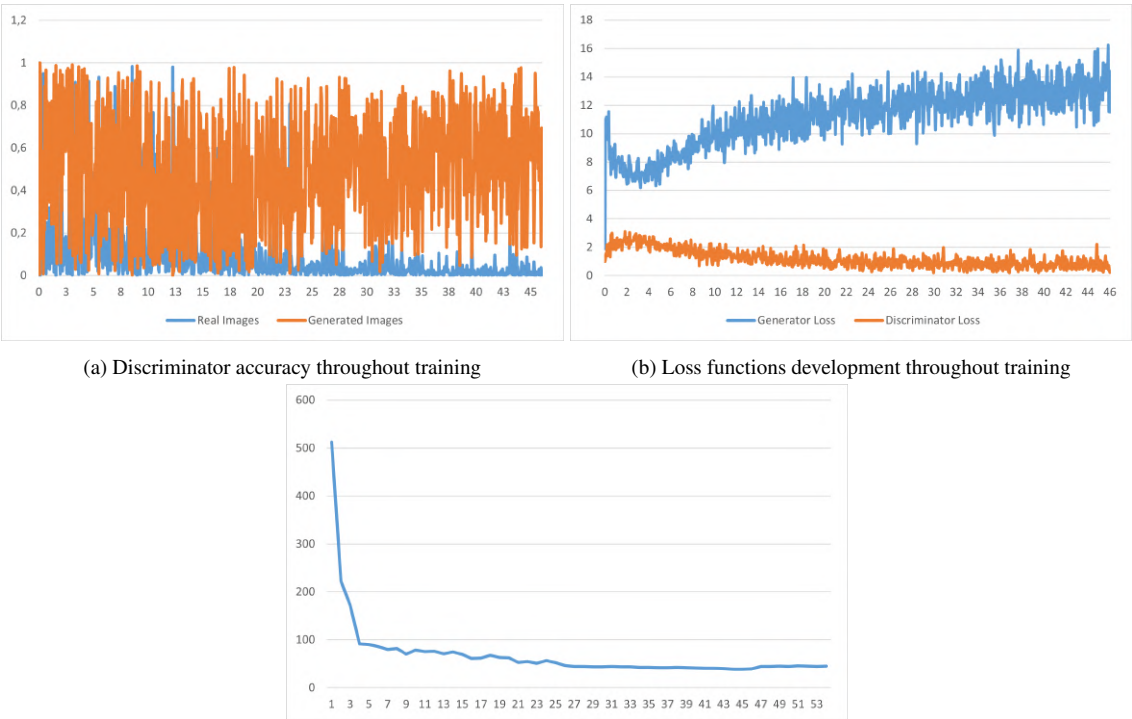
6.1.3 Loss Functions

Regarding the different loss functions used for training, the introduction of the dual contrastive loss did not result in a superior performance when compared to the original hinge loss function as shown in Table 6.3.

Test	IS	FID	KID
Hinge Loss 32 GSA	$\mathbf{2.109 \pm 0.034}$	**17.83**	$\mathbf{0.01211 \pm 0.00072}$
Dual Contrastive Loss	2.102 ± 0.028	39.30	0.04182 ± 0.00157

Table 6.3: Loss-related experimental tests and respective quantitative metric results throughout the development of the LWGAN for artificial CXR image generation. Bold values in each column indicate the best result for each metric.

Figure 6.5: Artificially generated CXRs from the *Dual Contrastive Loss* trained model



(a) Discriminator accuracy throughout training

(b) Loss functions development throughout training

Figure 6.6: Discriminator accuracy at predicting the image source (left), loss functions (right) and FID (center) values from the *Dual Contrastive Loss* model throughout training

Supporting the metric evaluation results, the overall image quality lacked better detail and a more correct representation of the anatomical structures, such as the vertebrae and ribs, than the models trained with the hinge loss, as shown in Figure 6.5. Additionally, the loss functions' evolution throughout training showed a rapid development from the start, the generator loss quickly

degraded leading to a gradually worsening model, as shown in Figure 6.6. The FID showed a rapid evolution until it started to stabilize and reach a plateau. As for the accuracy of the discriminator, it shows an irregular accuracy regarding the generated images, whereas the accuracy of the real image predictions showed a an irregular behaviour at the beginning of training and kept decreasing to zero throughout training.

6.1.4 Large Resolution

Figure 6.7 shows two generated samples by the *1024 Resolution* model. This model was the worst performing GSA-related model, since it did not converge and failed to train correctly. Table 6.4 compares this model's performance with the *32 GSA* model with which it is directly comparable, since both have the same hyper-parameters and only differ in resolution.

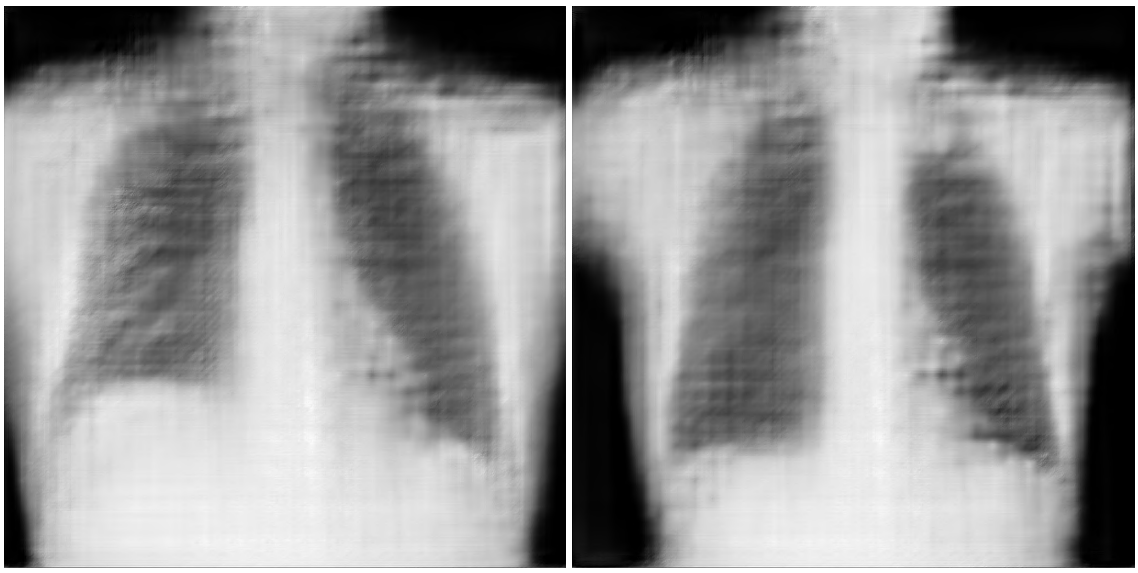


Figure 6.7: Artificially generated CXRs from the *1024 Resolution* model

Figure 6.8 presents the discriminator's accuracy at predicting the source of the images, the loss functions and the calculated FID for the *1024 Resolution* model. The three graphs show a normal behaviour and development throughout training, except the FID that does not improve over time.

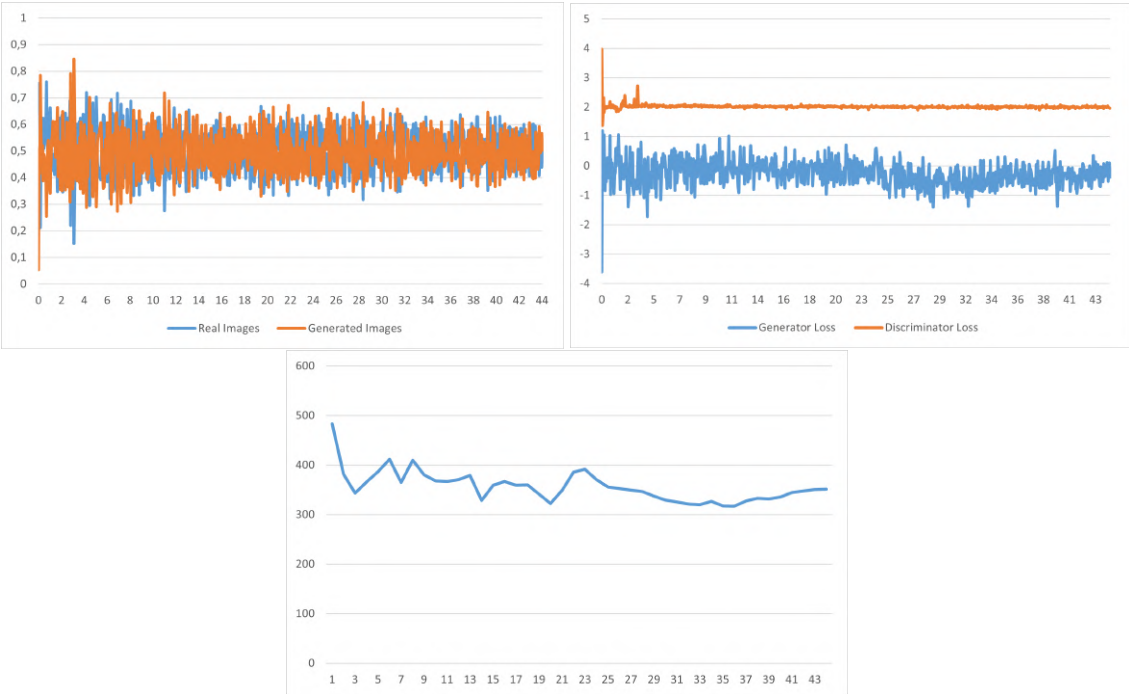


Figure 6.8: Discriminator accuracy at predicting the image source (left), loss functions (right) and FID (center) values from the *1024 Resolution* model throughout training

Test	IS	FID	KID
32 GSA 512×512	$\mathbf{2.109 \pm 0.034}$	**17.83**	$\mathbf{0.01211 \pm 0.00072}$
1024 Resolution 1024×1024	2.039 ± 0.031	316.03	0.44153 ± 0.00316

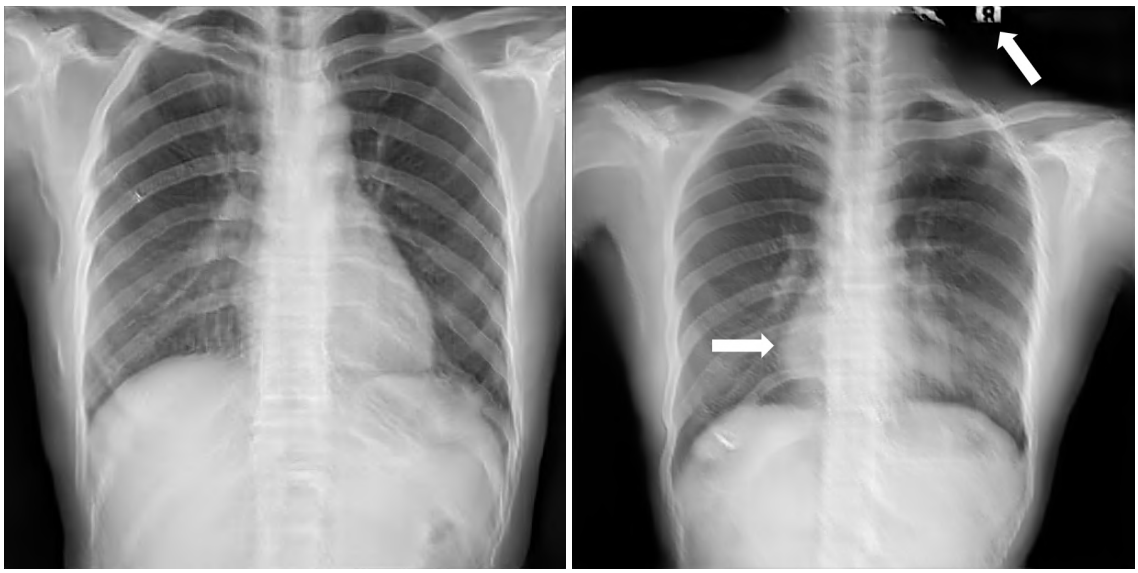
Table 6.4: Comparison between larger resolution and smaller resolution tests with GSA in the 32×32 layer and respective quantitative metric results throughout the development of the LWGAN for artificial CXR image generation. Bold values in each column indicate the best result for each metric.

6.1.5 Hyper-parameters

Finally, none of the tests performed for the assessment of the importance of the hyper-parameters selection outperformed the model trained with the 2000 image set and a batch size of 6 images. The results in Table 6.5 also show there was no improvement with an increased batch size or either an increased or decreased size of the image set. Additionally, the increased horizontal flip probability also leads to a poorer performance.

Test	IS	FID	KID
32 GSA	$\mathbf{2.109 \pm 0.034}$	**17.83**	$\mathbf{0.01211 \pm 0.00072}$
1000 Image Set	1.988 ± 0.043	94.80	0.11675 ± 0.00234
4000 Image Set	2.079 ± 0.042	50.96	0.05808 ± 0.00178
4000 Image Set Larger Batch Size	2.016 ± 0.031	57.78	0.0696 ± 0.00176
4000 Image Set Horizontal Flip $p = 0.5$	2.017 ± 0.036	60.74	0.07365 ± 0.00194

Table 6.5: Hyper-parameter related experimental tests and respective quantitative metric results throughout the development of the LWGAN for artificial CXR image generation. Bold values in each column indicate the best result for each metric.



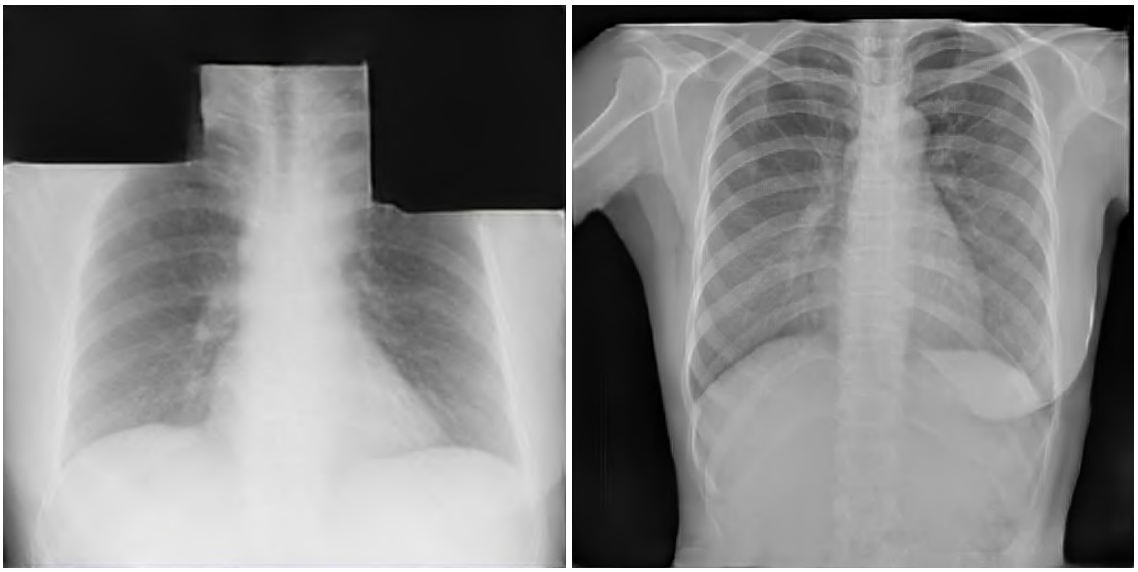
(a) *4000 Image Set* model

(b) Generated sample from the *4000 Image Set* model trained with the horizontal flip feature at $p = 0.5$, as shown by the heart on the left side of the image (right side of the patient) and inverted laterality marker. White arrows show horizontally flipped structures

Figure 6.9: Artificially generated CXRs from different Hyper-parameter related models

Figure 6.9 shows a generated sample from the *4000 Image Set* trained model, on the left, and a generated sample from the test related to the horizontal flip feature. The image on the left shows there is no particular improvement in the anatomical structure representation and overall detail.

As for the image on the right, it shows an image horizontally symmetrical to the real CXRs, as shown by the heart on the left side of the image (right side of the patient) and inverted laterality marker.



(a) Augmentation leak (b) One breast missing

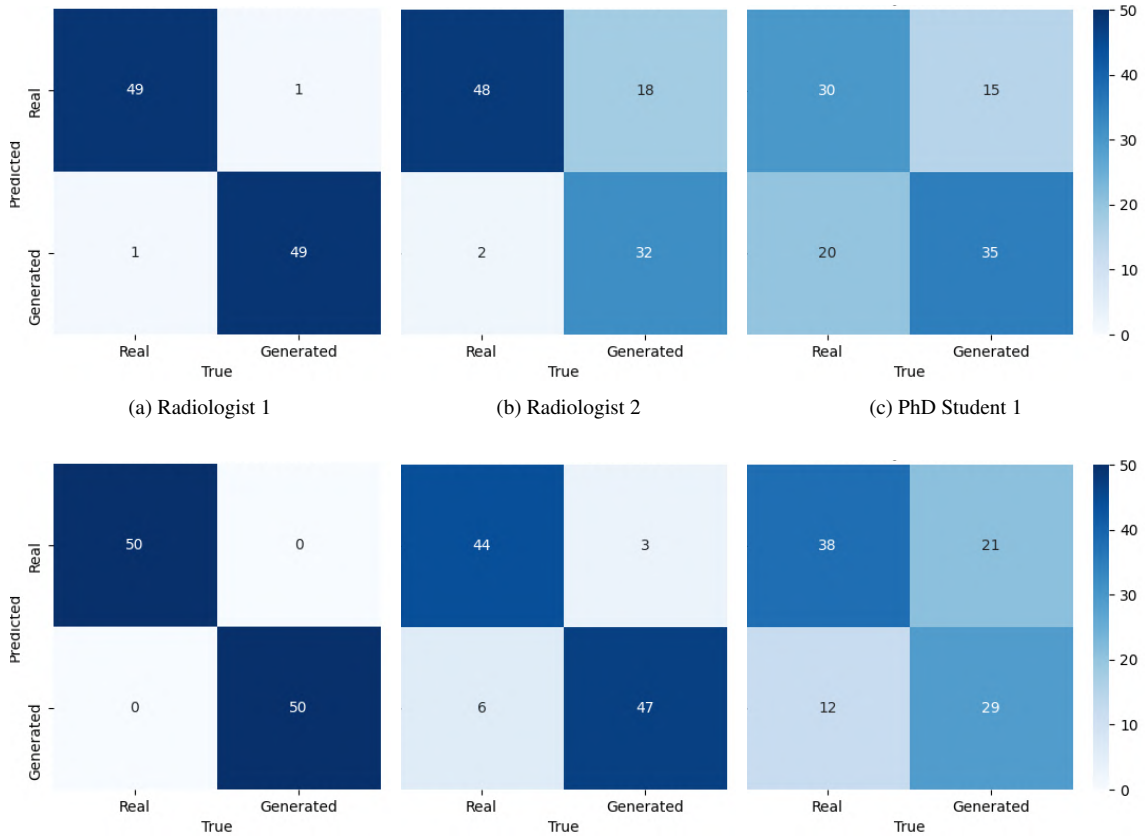
Figure 6.10: Examples of training failures from the trained models throughout the development of the GAN

Figure 6.10 presents two examples of failures during the training of the models. The image on the left shows an example of augmentation leaks during training. Training images are augmented to increase the number of training samples and in the case of a high probability for the augmentation operations, leaks of these augmentation samples can be learned by the model. In this case, the image shows two black rectangular cutouts in the two top corners. The image on the right shows another commonly found failed sample. In this image, only one breast can be found. Although in reality these cases do exist due to illness or trauma for example, there are not enough representations of those cases in the datasets for the model to learn it and represent it at such a large scale in the generated images.

6.2 Perceptual Validation

6.2.1 Authenticity Classification

The results of the image authenticity validation are shown in Figure 6.11, corresponding to the classification made by the radiologists, PhD students and the inexperienced individuals. Table 6.6 shows the accuracy, specificity and sensitivity regarding the classification by each participant. The specificity, or true negative rate, shows the number of images correctly classified as real in all truly real images in the test set. As for the sensitivity, or true positive rate, it shows the number of images correctly classified as generated in all truly generated images.



(a) Radiologist 1 (b) Radiologist 2 (c) PhD Student 1

(d) PhD Student 2 (e) Inexperienced Individual 1 (f) Inexperienced Individual 2

Figure 6.11: Authenticity classification by all six participants

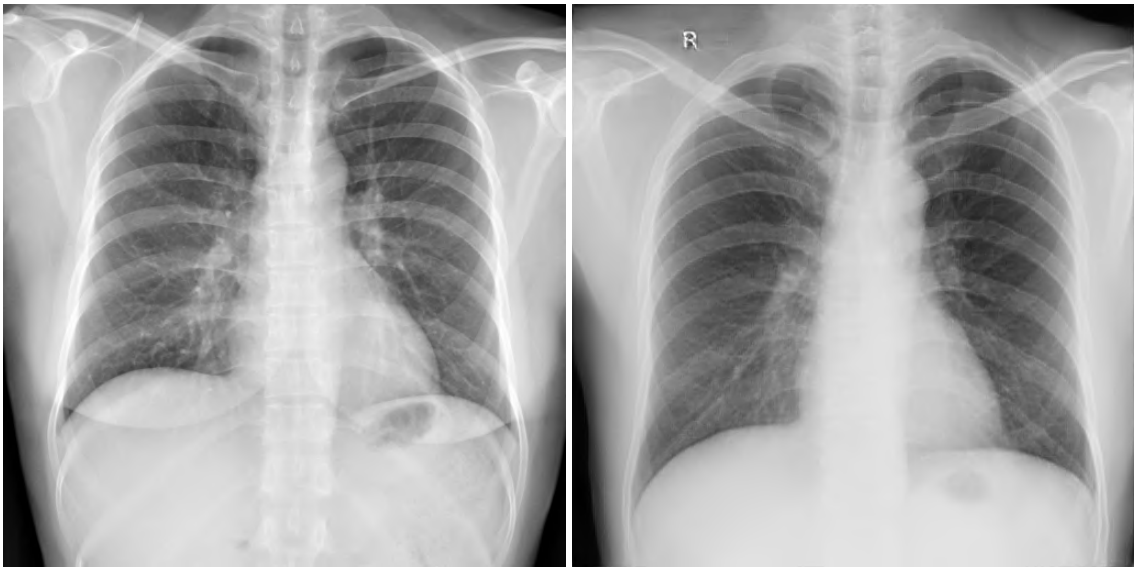
As can be observed in Table 6.6, the pair of radiologists performed the best at the overall classification task, followed by the PhD students and lastly by the inexperienced participants. As for the individual sets of images, the radiologists showed better accuracy at identifying real images than the other two pairs of participants, this time followed by the inexperienced individuals and then the PhD students. Regarding the sensitivity, the PhD students outperformed the two pairs, followed by the radiologists.

The PhD students were more capable at identifying the artificially generated images and the radiologists, as mentioned, the real images. The inexperienced participants did not outperform the two pairs in any of the image sources.

Participant	Accuracy	Specificity	Sensitivity
Radiologist 1	98/100 (98)	49/50 (98)	49/50 (98)
Radiologist 2	80/100 (80)	48/50 (96)	32/50 (64)
PhD student 1	65/100 (65)	30/50 (60)	35/50 (70)
PhD student 2	100/100 (100)	50/50 (100)	50/50 (100)
Inexperienced individual 1	91/100 (91)	44/50 (88)	47/50 (94)
Inexperienced individual 2	67/100 (67)	38/50 (76)	29/50 (58)

Table 6.6: Authenticity classification accuracy by the two Radiologists. Values in parenthesis are percentages.

Figure 6.12 presents one sample of each image source, real and generated, used in the classification task.



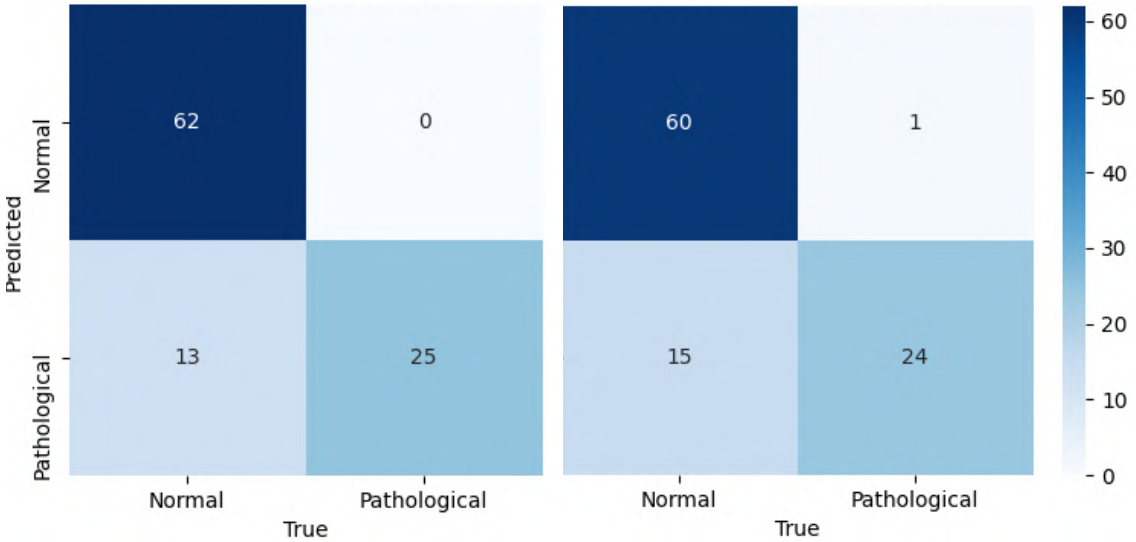
(a) Real normal image (b) Artificially generated normal image

Figure 6.12: Representation samples of both real and generated images used in the authenticity classification task

6.2.2 Normality Classification

As mentioned in Chapter 5, the two radiologists performed the additional task of classifying each of the 100 images as normal or pathological. Figure 6.13 shows the confusion matrix with the results for the normality classification. Both radiologists achieved similar results, having classified most of the pathological images with the correct label corresponding to high sensitivity. However, both radiologists misclassified normal images, leading to a lower specificity.

Table 6.7 shows the results for the perceptual validation of each type of image. To evaluate the performance of the radiologists in the normality classification, the images were divided into real and artificial, which allows for an interpretation of the structural representation in each type



(a) Radiologist 1 (b) Radiologist 2

Figure 6.13: Label classification of 100 test images by both Radiologists

of image. Both radiologists performed poorly on the classification of real normal images, which means a significant number of these images were classified as pathological.

	Rad 1	Rad 2	Average
Accuracy	87/100 (87)	84/100 (84)	85.5/100 (85.5)
Specificity	62/75 (83)	60/75 (80)	61/75 (81.4)
Sensitivity	25/25 (100)	24/25 (96)	24.5/25 (98)
Specificity Real	16/25 (64)	18/25 (72)	17/25 (68)
Specificity Artificial	46/50 (92)	42/50 (84)	44/50 (88)

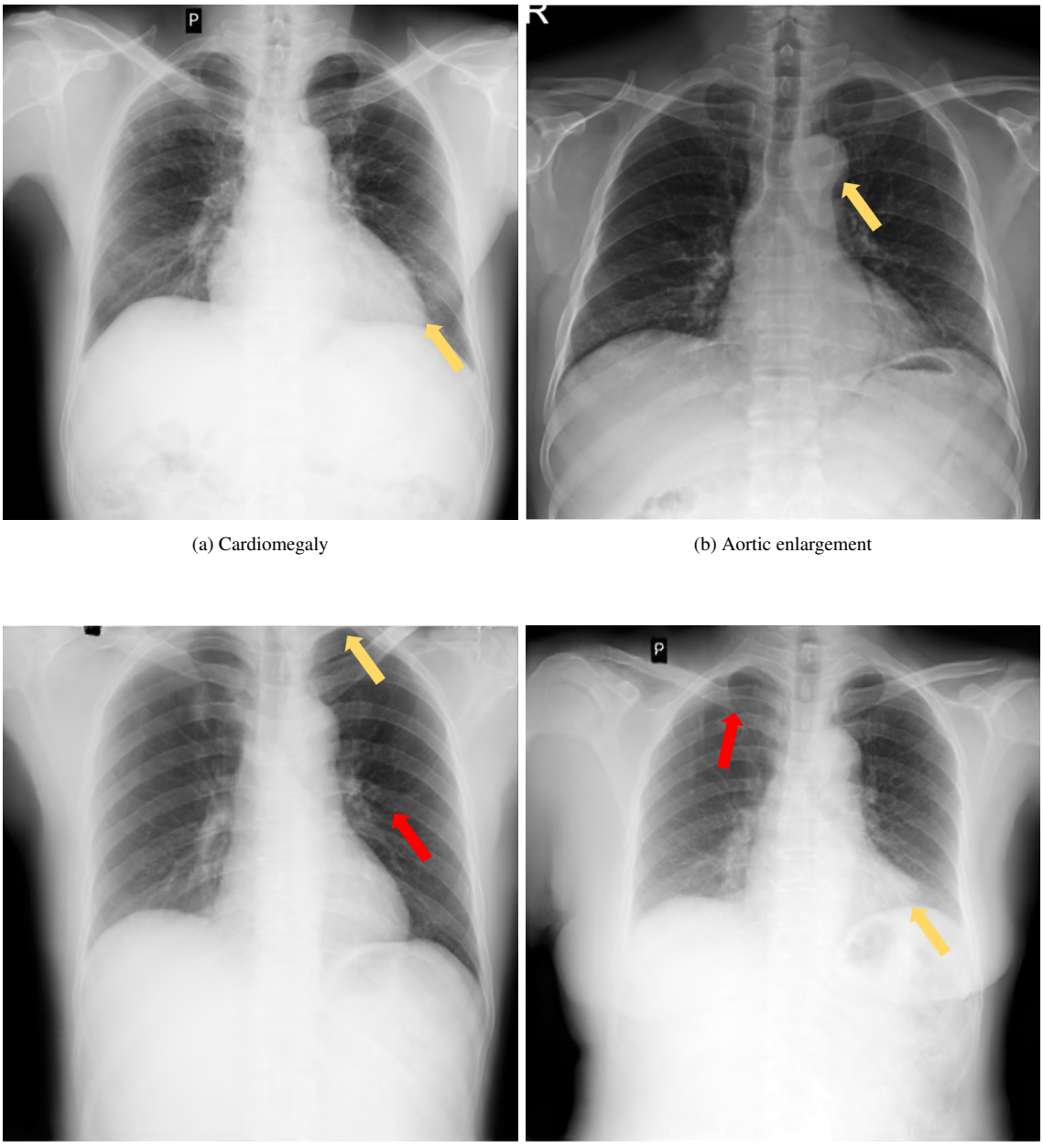
Table 6.7: Normality classification accuracy by the two Radiologists. Values in parenthesis are percentages.

Figure 6.14 shows four images used in the authenticity and normality classification tasks. The two images on the top row are real and were annotated as normal by the responsible group of radiologists. However, the two radiologists that participated in the classification tasks deemed them incorrectly labeled. The image on the top left shows a cardiomegaly, as indicated by the arrow. The image on the right shows a CXR with an aortic enlargement. Both of these findings were not identified by the VinDr-CXR radiologists and were labeled as normal.

The two images in the bottom row were artificially generated and classified as such and also as pathological due to structural irregularities. The image on the bottom left shows an apical asymmetry, which is what led to it being classified as pathological by the pair of radiologists. Additionally, a distorted clavicle was mentioned by one of the radiologists as a the structural irregularity that led to it being classified as an artificial sample.

The image on the right was classified as pathological due to a cardiomegaly. Additionally, the

radiologists were led to believe it was fake due to an irregularity in the clavicle, as shown by the arrow.



(a) Cardiomegaly

(b) Aortic enlargement

(c) Apical asymmetry

(d) Incomplete clavicle

Figure 6.14: Two examples of real normal images classified by the two participating radiologists as pathological (top) and two artificially generated normal images correctly classified as fake but incorrectly classified as pathological by the radiologists due to structural irregularities. Yellow arrows show the pathological finding and red arrows the structural irregularities

Summing up the results from the authenticity classification, Table 6.6 pairs the accuracy achieved by each group of participants. The overall best performance was achieved by the pair of Radiologists.

6.3 Application in Image Classification/Detection

6.3.1 Binary Classification Model

Figure 6.15 shows the two relative frequency histograms of the predicted abnormality probability by the MobileNet architecture for the real and artificial images.

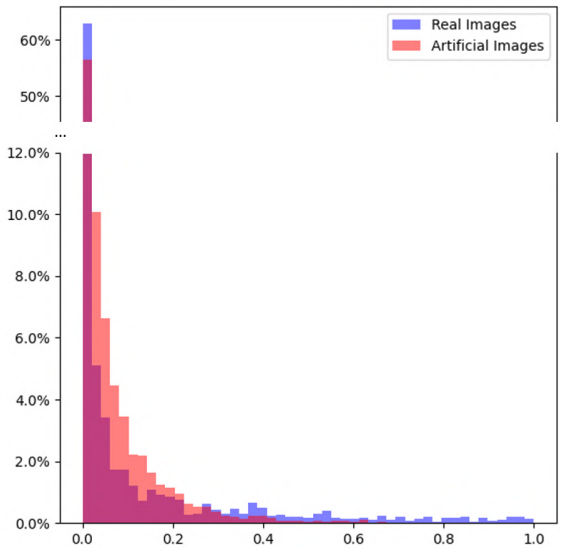


Figure 6.15: Relative frequency histogram for the real and generated images of binary abnormality classification of CXR images

It can be seen that the set of real images has a higher percentage of very low probability images, corresponding to a prediction of normality. However, the real image set also has a higher percentage of higher probability predictions, corresponding to a prediction of pathology.

Image Set	Classification	
	Normal	Pathological
	Total (%)	Total (%)
Real Images	1979 (93.35)	141 (6.65)
Artificial Images	5947 (99.12)	53 (0.88)

Table 6.8: Binary classification accuracy on the real and artificial images.

Table 6.8 shows the accuracy of the binary classification model with regard to each set of images. The overall results were similar, with a slight performance increase in the artificial image dataset.

6.3.2 Training of a Pathology Classifier

To assess the change in performance of each model, the classification accuracy of each class was evaluated, along with the average of each model's accuracy. All of the pathological classes were

classified with the trained YOLOv5 network.

Class	A	B	C
Aortic Enlargement	0.55	0.61	**0.87**
Atelectasis	**0.30**	0.22	0.27
Calcification	0.09	0.09	**0.11**
Cardiomegaly	0.56	0.68	**0.87**
Consolidation	0.19	**0.27**	0.23
ILD	0.26	0.24	**0.28**
Infiltration	0.21	0.26	**0.28**
Lung Opacity	0.16	0.24	**0.27**
Nodule/Mass	0.22	0.28	**0.33**
Other Lesions	0.09	0.11	**0.14**
Pleural Effusion	0.50	0.51	**0.54**
Pleural Thickening	0.15	0.17	**0.24**
Pneumothorax	0.34	**0.58**	0.54
Pulmonary Fibrosis	0.30	0.35	**0.38**
Average	0.33	0.37	**0.42**

Table 6.9: AUC of the precision-recall curve of each trained model for each pathology class. Column A represents the model trained with real pathological images, column B represents the model trained with real pathological and artificial normal images and column C represents the model trained with real pathological and normal images.

Table 6.9 shows the area under the precision-recall curve of each model for each of the pathological classes with the best-performing model being highlighted, and Figure **??** shows a graphical representation of the same comparison. The best performing model was the one trained with both real normal and pathological images, followed by the one trained with real pathological images and artificial normal images, and lastly the model trained with just real pathological images.

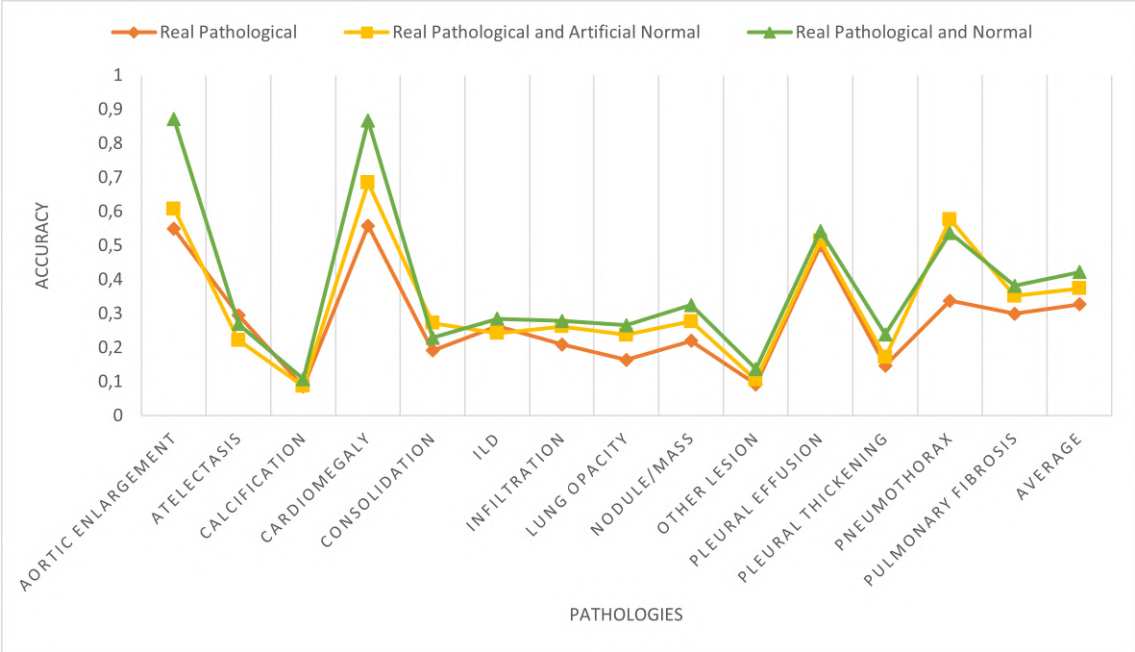


Figure 6.16: Graph showing the comparison between the area under the precision-recall curve of each model for each of the pathological classes

Chapter 7

Discussion

In this section, the results of the LWGAN evaluation will be analysed and interpreted. Overall, the results in the tables in the previous chapter indicate that the overall best performing architecture is the one described by the *32 GSA* test with an image resolution of 512×512 pixels, self-attention in the 32×32 layer, 2000 training images, a batch of 6 images, augmentation probability at 0.25 and translation, cutout and color as augmentation techniques. It showed the best performance regarding the evaluation metrics and also the overall image quality.

7.1 Chest X-Ray Image Generation

7.1.1 Quantitative Metric Evaluation

There is no publicly available related work to compare with the results of the evaluation metrics used to assess the performance of this GAN, therefore, the performance is comparable only within the performed experiments of the LWGAN.

The FID is a biased metric, dependent on the number of evaluation images, the IS is suboptimal and does not take into account the intra-class variability of the images, and the KID has a high variance. Although the three evaluation metrics are flawed, these are considered to be the state-of-the-art evaluation metrics for GANs. However, when combined, the three metrics have shown to adequately represent the quality of the models and contribute for the overall interpretation of a GAN's performance.

Through the comparison of the generated images and the evaluation metric results used to compare the experimental tests it is possible to conclude that the evaluation metrics do adequately represent the quality of the images when used together. The FID, as shown in the tables of the previous chapter, has a strong correlation with the values of the KID. However, the same does not always apply for the IS when compared to the other two metrics. The IS value for the *256 GSA* test, for example, does not translate the global performance of the generated images.

7.1.2 Resolution

The first experiment was related to the image resolution. Although the easiest model to train would be the one trained with 256×256 resolution images, the best performing model turned out to be the one trained with the 2000 images of 512×512 pixels. The number of representative features in larger resolution images should have made it harder to train, nevertheless, it showed a larger ease of training and better results. Additionally, the training time was similar to the training time of the lower resolution model.

7.1.3 GSA

The GSA implemented in the second experiment showed a significant improvement on the overall performance of the LWGAN. As mentioned in the previous chapter, the best performing model had GSA in the 32×32 layer. The comparison between the tests with GSA in different layer sizes shows that it is especially impactful in smaller layers, enabling the network to connect lower resolution long-distance features between them. When compared against the tests with GSA in larger layers or multiple layers, the end result of this operation led to a more detailed image with better representations of the anatomical structures. It is not clear why the models with GSA in multiple layers, including the 32×32 layer, experienced worse results to such an extent. Nevertheless, there is no indication that by combining layers with GSA, such as a good performing layer, 32×32, with a bad performing layer, 256×256, would improve the overall performance. It is also possible that this problem might be associated with difficulty in converging and not with the architecture in it self.

Regarding the loss functions and discriminator prediction accuracy of the *32 GSA* model, it can be seen that a collapse of the training occurred on epoch 32. During training, discriminator and generator are kept in a close balance, with a discriminator loss close to 2 and generator loss close to 0 indicating that the discriminator cannot successfully distinguish real from artificial images. However, Figure 6.3 shows a disruption of this balance and collapse of the GAN with a steep increase in the generator loss and a decrease in the discriminator loss, leading to a near perfect classification of both real and artificial samples (Figure 6.3). However, the model's performance in terms of FID (Figure 6.2) improved at a faster rate up until epoch 50 where it then began to worsen at a slower rate. From this point onward, it kept getting worse and never achieved the same performance again. With this in mind, the discriminator accuracy supported the idea of vanishing gradients, since the discriminator got too good at classifying the images, however, as mentioned, there was an improvement in the overall performance of the model.

Comparing the *32 GSA* model to those with GSA in larger layers, in particular to the *256 GSA* model for example, it can be seen that both quantitative and qualitative results were worse with GSA in larger layers. Shown in Figure 6.4 are two samples generated by the *256 GSA* model, with more examples shown in Figures A.3 and A.4. The overall pixel intensity of both images is largely different. The set of image tiles in Figure A.4 shows that this occurrence is common in most of the generated images by the *256 GSA* model. This model includes GSA in the 256×256 layer,

which is believed to have led to this difference in the generated images, since the occurrence did not happen with any of the remaining trained models.

With this in mind, comparing this model's performance with the *32 GSA* model, which includes GSA in the 32×32 layer, and other models with GSA in multiple layers, one can conclude that GSA in a larger resolution layer can likely lead to an increased focus on irrelevant features.

7.1.4 Loss Functions

The dual contrastive loss function did not outperform the model trained on the hinge loss function. However, the development of the generator and discriminator loss functions showed a promising development. As Figure 6.6 shows, the generator and discriminator loss functions adapted quite rapidly and at a faster pace than the hinge loss, however, after epoch 3, the generator's loss function output started increasing. Similarly to the losses of the *32 GSA* model, the *Dual Contrastive Loss* model did improve up until epoch 44, where it began to progressively worsen.

Regarding the generated image quality, although the evaluation metric results are amongst the best, the anatomical structures such as ribs and the distance between them, shoulder blades and vertebrae are poorly represented. Additionally, the overall image resolution and detail lack the sharpness required to properly identify structures and consider the images good artificial samples.

7.1.5 Large Resolution

Regarding the results of the *1024 Resolution* model, although the loss functions developed as expected, maintaining a stable and constant value for both the discriminator and generator network, which in case of this implementation of the LWGAN with the hinge loss is at around 2 and 0 approximately, the model did not converge and did not improve the overall generation quality. As observed in Figure 6.7, the FID did not improve significantly over the 43 epochs and stabilized. The fact that the accuracy and loss functions behaved as expected for a normal test, it shows that both the discriminator and generator failed to learn the representations and there was no collapse of the overall training.

Unfortunately, since it did not train correctly, higher resolution images could not be generated and the original 512×512 resolution was kept during the following experiments and GAN validation.

7.1.6 Hyper-parameters

The hyper-parameter related tests showed that the LWGAN is sensitive to variability in these parameters. However, neither decreasing or increasing the number of available images for training improved the overall performance of the model, when compared to the best trained model. Additionally, a larger batch size did not improve the performance as well, even when compared against a model trained with the same number of training images. In deep learning applications, an increase in the volume of training data leads to an improvement until a saturation point is reached where adding further samples does not lead to an increase in performance. However, in smaller

more compact variations, this saturation level may be reached with a smaller amount of images. The models trained with 4000 images did not improve in comparison to the models trained with 2000 images, however it is not clear if this is caused by the amount of training images or other aspect of training, such as the loss function behaviour of different models. Nevertheless, all models trained on 4000 images showed inferior performance when compared to the models trained with 2000 images.

Regarding the model trained with the horizontal flip augmentation with 0.5 probability, it was shown that the higher probability led to an augmentation leak, where the generator learns to generate not only the original image characteristics but also the introduced augmentation. In this case, this leads not only to a worse performance in terms of FID, KID and IS but also leads to the generation of images which are not suitable to be used in clinical practice. With many deep learning applications, especially health and medical applications such as CXRs, input from medical doctors, in this case radiologists, is essential. To be accepted for clinical use, deep learning algorithms have to be validated by someone with field-knowledge. Therefore training these models while empirically validating the results and methods with medical doctors is crucial.

The right side image in Figure 6.9, as mentioned in the previous chapter, shows an example of the augmentation leak of the horizontal flip. In this case, the heart and label marker are some examples introduced during training that can negatively affect the training of the model and are not suitable for clinical practice.

The two examples showed in Figure 6.10 occurred due to the model learning incorrect representations of the statistical representation of the data. In the case for the augmentation leak, having a large augmentation probability, which is the probability of a batch being augmented, can lead to a large number of augmented samples being fed to the model's input for training. These images do not correctly represent and, just as the images with a horizontal flip, are not suitable for use in clinical practice.

Nevertheless, as with the horizontal flip feature, data augmentation, when introduced in a smaller scale, does not negatively affect the training of the model, since the image volume is not large enough to represent those features as a statistical significance.

The image on the right shows only one breast, which in reality and clinical practice does exist. However, the volume of generated images with this structural difference is quite significant, which supports the idea that the model is not learning the statistical representation of the data. It is assumed that the model is mixing features from CXRs taken from male patients with features from CXRs taken from female patients. This operation is common in GANs, however, for the use in clinical practice it is not clear if it will introduce bias or other limitations to classification models.

7.1.7 Summary

Despite the results for the hyper-parameter related tests, or any of the above mentioned tests for that matter, it should be taken into account that none of the mentioned tests had a loss function behaviour similar to the one in the *32 GSA* model, except for the *512 Resolution* model (Appendix A.14). Since the larger performance improvement in the *32 GSA* model originated simultaneously

to a collapse of the discriminator-generator balance, it is not entirely correct to assume that the performance of the remaining models with different loss function behaviour would underperform when compared to a *32 GSA* model if the same imbalance had occurred. Figure 6.2 shows that around epoch 33, before the FID curve dove to a better performance, the FID was at roughly around 63. It is still hard to form any conclusions, however, it shows that the performance of all other models would likely not be far behind the best performing model, if the same GAN collapse and simultaneous improvement had occurred.

Regarding the overall model performance, none of the trained models showed indications of mode collapse, i.e., did not constantly produce the same plausible output that could be misinterpreted by the discriminator. Every model showed a varied set of images without obvious repetition, as shown in the mosaic figures in the appendix A.

Nevertheless, the best overall performing LWGAN model is the *32 GSA* model, which is capable of generating high-quality CXR images with adequate detail, correctly represented anatomical structures and without suffering from mode collapse or other limitations associated with GANs.

7.2 Perceptual Validation

7.2.1 Authenticity Classification

The six participants of the authenticity classification task presented interesting results. After questioning the participants and asking the reason that led to the classification choices, two main approaches were described: three participants noticed the difference in detail and resolution of the structures and based their decisions on that parameter, and the remaining three participants tried to find structural irregularities and did not take into account the detail and resolution as much as the previous three. The main irregularities sought after by the participants were details such as ripples in the edges of the bones, distortion of the clavicles, asymmetries and angulation of other anatomical structures, with some examples shown in Figure 6.14.

Radiologist 1, PhD Student 2 and Inexperienced Individual 1 all followed the first approach. As expected, since the generated images are not perfect with regard to detail and pixel resolution, all three performed very well, with an overall average accuracy of over 96%.

The other three participants, which took into account the structural representation and overall quality of the generated image, were not as accurate at distinguishing real from generated images. Radiologist 2 mentioned that some of the generated images posed as lower quality real CXRs, which are quite common in reality. With this in mind, the decision process of radiologist 2 was not influenced by this issue and since the structural elements were correctly represented, generated images looked real and were classified as so. The same logic was proposed by the two other participants. The overall average accuracy in the authenticity classification by these three participants was above 70%, quite different from the first three participants.

As for the split into real and artificial images, specificity and sensitivity, as expected, real images were more often correctly classified as real images than the generated images. This means

that real images were easier to identify and generated images were often classified as real, which in an ideal situation would happen with every image.

The best performing pair of participants was the pair of radiologists, followed by the PhD students and lastly the inexperienced individuals. One may conclude that higher level representations and overall image quality are enough to lead to incorrect classifications for the pair with less field-knowledge, and as it increases, it becomes easier to notice the differences and main limitations of the generated images.

7.2.2 Normality Classification

The results of the normality classification task performed by the two radiologists were not as originally expected. Both radiologists accurately labeled the pathological images, with an average accuracy of 98%, as expected. However, for both individuals, the labeling of the normal images was not accurate, having both incorrectly labeled a similar amount of images. Upon a more detailed review of the image labeling, it was concluded that a significant amount of the incorrectly labeled images by both radiologists were the same, i.e., both radiologists incorrectly labeled the same images.

When discussed with the two radiologists, they confirmed that most of these images could not be considered normal and that the ground truth label was incorrect, meaning that images labeled as normal on the dataset presented in fact pathologies which were not identified/labeled by any of the radiologists that participated in the original annotation of the dataset. This meant that from the 25 real normal images, a significant amount (up to 7 images) was mislabeled and the images were in fact pathological images.

These cases occurred more commonly in real images than in the generated images, where only 6 images in a total of 50 were classified as having pathologies. Additionally, the findings found in the generated images that led to the two participating radiologists classifying them as pathological, were mainly distortions in anatomical structures, whereas with real images, these findings were identifiable pathologies belonging to the VinDr-CXR dataset classes. The most common pathologies identified by the radiologists in real images were cardiomegaly as well as opacities/nodules in the ribs and clavicles. Aortic enlargement and enlarged mediastinum were also identified.

Since the VinDr-CXR dataset is the only large volume dataset with manual annotations, it is important to be aware of the large amount of misannotated labels. This dataset may be preferable when compared with other large datasets that are not manually annotated, however, it has shown to be flawed and future studies should have this into account.

7.3 Application in Image Classification/Detection

7.3.1 Binary Classification Model

The results shown in Table 6.8 show a larger number of artificial images were classified as normal images by the binary classifier.

As discussed in the previous section, a small but significant amount of images in the VinDr-CXR dataset appear to be incorrectly labeled. These images are mainly labeled as normal images by the radiologists who performed the original dataset annotation.

The influence of the mislabeled images can be seen in the histogram of classification predictions for real images, as shown in Figure 6.15, where the classification model classifies some of the real images as abnormal with a high level of certainty. Since the LWGAN learns a statistical representation of the features of the images, the model does not learn to generate abnormal images as normal even though they may be present in the training dataset, since these have significantly different features and are likely considered as outliers during training.

Figure 6.15 shows a higher number of real images classified as normal images with a high level of certainty. However, it also classified a large number of images as abnormal. As for the generated images, these are not classified as normal with such a high level of certainty, since the LWGAN is not capable of generating perfect authentically-looking CXRs. Nevertheless, a larger number of artificial images are classified as normal, shown as light red in the figure, and very few images are classified as abnormal.

The results support the idea of the VinDr-CXR dataset being a flawed dataset. Additionally, the generated images seem to accurately represent and pose as real normal CXRs when analysed by an objective classifier with regard to the overall image structure. Directly comparing the distributions of the real and artificial images one might even consider that the artificial images are more normal than the normal data from which they were generated, since the pathological outliers have been filtered out. This is supported also by the perceptual validation performed by the radiologists and the higher specificity obtained in artificial images in normality classification.

7.3.2 Training of a Pathology Classifier

The use of artificial images for the training of classification models is one of the goals of image generation. In an ideal scenario, the addition of artificially generated images would provide the same training performance as if the model had been trained with the same number of real images.

As mentioned in the previous chapter, Table 6.9 and Figure **??** show the performance of three trained classifiers. The goal was for the model trained with real pathological images and artificial normal images to be able to outperform the model trained with real pathological images, which it did successfully. Ideally, the generated images would be on par with real normal images and be as successful at training a classification model. However, and as expected, since the LWGAN is not capable of generating images that completely represent and pose as real images, the result of the

classifier trained on artificial images did not achieve the same performance as the one trained only on real images.

Nevertheless, as mentioned, when compared to the performance of the model trained with only real pathological images, the model trained on both real pathological and artificial normal images has shown to improve the overall classification performance. This indicates that, although the generated images are not on par with real images, they still provide an improvement on the overall performance of the classification model for pathological classes.

7.4 Limitations and Future Work

The work developed throughout this dissertation met several constraints. The main limitation found in the whole development stage was the dataset. The VinDr-CXR dataset is the only publicly available completely manually labeled CXR dataset, which means it is not susceptible to NLP annotation errors like other commonly used datasets, giving it an advantage when compared to those other datasets. Nevertheless, it turned out to have a significant amount of incorrectly labeled images. An incorrectly annotated dataset leads to poorly trained models and is a major disadvantage for future clinical applications. High quality datasets are one of the main limitations for CXR automatic diagnostics systems, since all of the available datasets, including the VinDr-CXR, are flawed. In future work and development, an improved or new dataset should be a priority, for it impacts the overall performance of the models, especially in applications where only small amounts of data are available.

Throughout the development of the LWGAN models, a large set of data was available, since the *No Finding* class holds the majority of the available images. However, training with only the *No Finding* class is a limitation since it does not train models on pathological findings, which are less represented classes. Additionally, most cases who require some degree of explainability of the model's decisions are cases with pathological findings. Some of the available pathological images in the VinDr-CXR dataset have more than one pathology, which can lead to conflicts during training. Ideally, each training image would contain findings related to just one pathological class. This would allow the models to train and learn specific features to each single pathology. In future developments there should be a focus on having models trained on pathological images, since it is where the data representation limitation lies. Another aspect that could be advantageous is the idea of the disentanglement of features, which could allow specific pathologies to be added to specific locations in normal images, such as nodules. Parallel to this, human perceptual validation should be included in the development, since it is essential for a correct training and progression of the models to be later used in clinical applications.

A major limitation of the work developed in this dissertation is the quality of the generated images. Although the 512×512 resolution appears to be enough for training classification models, the goal is to fully represent the native data, even regarding the resolution. With this in mind, the work developed in the future should have a focus in progressively growing a few-shot GAN, in

this case the LWGAN, to allow for larger resolution images to be generated. The available data has the required quality for it and should be one of the main focuses.

GAN evaluation metrics are one of the commonly known limitations regarding this type of algorithm. Although the evaluation metrics used in this work were capable of providing additional information regarding the LWGAN's performance, none of these metrics adequately represents the whole range of quantitative analysis that is required to understand the performance and quality of a GAN. This is an issue that is actively addressed by the GAN development community and is a problem with a much needed solution.

These limitations show that the work developed in this thesis is a stepping stone to future developments. These limitations can be looked upon and new methodologies and tools can be developed in future work to further improve the algorithms.

Chapter 8

Conclusion

The state of the art of deep learning models for automated pathology classification systems has been progressively improving, but significant limitations remain due to the available data and the lack of knowledge regarding the decisions made by these models.

The goal of this dissertation was to research and develop the artificial generation of high quality CXR images when trained on small datasets. The final purpose is to be able to provide medical imaging and computer vision researchers with better tools to allow for more robust methods of automatic detection or diagnosis of several pathologies. The artificial generation of realistic CXR images would allow to supplement the amount of available training data for automatic multi-label diagnostic systems and to provide some degree of explainability of the decisions made by deep learning models. The process began by understanding the necessity regarding diagnostics of CXR images and automatic chest radiography analysis, followed by reviewing the best available datasets. Secondly, the state of the art of image generation and evaluation was reviewed and researched, with a focus on generative adversarial networks and commonly used evaluation metrics.

This thesis' contributions lie on the LWGAN development methodology to explore the potential of the architecture and evaluation methods to successfully assess its performance. Regarding the architecture, a systematic approach was explored, where the architecture was empirically developed with regard to several key aspects, including image resolution, global self-attention, loss functions and the model's hyper-parameters. The developed models achieved an adequate performance, generating high quality images capable of being used in the training of classification models while improving the overall performance.

To ensure a complete validation of the developed solutions, both quantitative and qualitative methods of evaluation were employed. The quantitative evaluation was ensured by the FID, KID and IS. The results showed that these metrics, when combined, complement each other and provide a useful quantitative evaluation of GANs. Nevertheless, these metrics are known to have limitations. The qualitative evaluation was assessed by a perceptual validation performed by several participants with different experience levels in CXR image analysis, who showed that the generated samples still lack the detail and resolution of real images, however, represent the anatomical structure of real CXRs. Finally, the models were also assessed by the performance on two applica-

tions in image classification and pathology detection. The results in both applications showed that the generated images are high quality images with a correct representation of the native CXRs, adequate to be used in classification models and successfully improve overall classification performance in pathological classes.

Altogether, the development of this work was essential for an understanding of the needs and limitations of generative models in medical image applications. It was done in a structured manner with a complete assessment of the models' performance. The used architecture, the LWGAN, enabled a fast development due to its short training time and low computational needs, and has shown to be efficient for the development of a quality GAN that is able to generate quality images.

References

[1] Darcy D. Marciniuk, Dean E. Schraufnagel, Thomas Ferkol, Kwun M. Fong, Guy Joos, Victorina López Varela, and Heather Zar. Forum of International Respiratory Societies The Global Impact of Respiratory Disease. Technical report, WHO, 2012.

[2] Julio Mendoza and Helio Pedrini. Detection and classification of lung nodules in chest X-ray images using deep convolutional neural networks. *Computational Intelligence*, 36(2):370–401, may 2020.

[3] Yu Xing Tang, You Bao Tang, Yifan Peng, Ke Yan, Mohammadhadi Bagheri, Bernadette A. Redd, Catherine J. Brandon, Zhiyong Lu, Mei Han, Jing Xiao, and Ronald M. Summers. Automated abnormality classification of chest radiographs using deep convolutional neural networks. *npj Digital Medicine*, 3(1), dec 2020.

[4] Jared A. Dunnmon, Darvin Yi, Curtis P. Langlotz, Christopher Ré, Daniel L. Rubin, and Matthew P. Lungren. Assessment of convolutional neural networks for automated classification of chest radiographs. *Radiology*, 290(3):537–544, mar 2019.

[5] Yifan Mao, Fei Fei Xue, Ruixuan Wang, Jianguo Zhang, Wei Shi Zheng, and Hongmei Liu. Abnormality Detection in Chest X-Ray Images Using Uncertainty Prediction Autoencoders. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12266 LNCS:529–538, oct 2020.

[6] Nina Tuluptceva, Bart Bakker, Irina Fedulova, Heinrich Schulz, and Dmitry V. Dylov. Anomaly Detection with Deep Perceptual Autoencoders. *arXiv*, jun 2020.

[7] R. F. Mould. The early history of x-ray diagnosis with emphasis on the contributions of physics 1895-1915, 1995.

[8] G. J. Bansal. Digital radiography. A comparison with modern conventional imaging. *Postgraduate Medical Journal*, 82(969):425–428, 2006.

[9] Mikael Häggström. X-ray on the left side of a 31 year old male, showing normal anatomy and no injuries.

[10] Martin Hoheisel. Review of medical imaging with emphasis on X-ray detectors. *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 563(1):215–224, 2006.

[11] Portable versus Fixed X-ray Equipment: A Review of the Clinical Effectiveness, Cost-effectiveness, and Guidelines - NCBI Bookshelf, 2016.

[12] European Commission. Radiaton Protection N° 180 Medical Radiation Exposure of the European Population (Part 1/2). Technical report, European Comission, 2015.

[13] Computed Tomography (CT).

[14] Pei Jan Paul Lin. Technical advances of interventional fluoroscopy and flat panel image receptor. In *Health Physics*, pages 650–657. Health Phys, nov 2008.

[15] Juan José Vaquero and Paul Kinahan. Positron Emission Tomography: Current Challenges and Opportunities for Technological Advances in Clinical and Preclinical Imaging Systems. *Annual Review of Biomedical Engineering*, 17(1):385–414, dec 2015.

[16] Vijay P.B. Grover, Joshua M. Tognarelli, Mary M.E. Crossey, I. Jane Cox, Simon D. Taylor-Robinson, and Mark J.W. McPhail. Magnetic Resonance Imaging: Principles and Techniques: Lessons for Clinicians. *Journal of Clinical and Experimental Hepatology*, 5(3):246–255, 2015.

[17] Michigan Medicine and University of Michigan. Pleural Diseases | Michigan Medicine.

[18] Heart Anatomy | Anatomy and Physiology II.

[19] B Natrajan, M Prabakaran, and B Natrajan. Cardiomegaly. In *Diagnostic Atlas of Pediatric Imaging: Chest and Mediastinum*, pages 88–88. Jaypee Brothers Medical Publishers (P) Ltd., nov 2010.

[20] Shaney L Barratt, Andrew Creamer, Conal Hayton, and Nazia Chaudhuri. Clinical Medicine Idiopathic Pulmonary Fibrosis (IPF): An Overview. *Journal of Clinical Medicine*, 2018.

[21] Ayla Al Kabbani and Yuranga Weerakkody. Lung atelectasis | Radiology Reference Article | Radiopaedia.org.

[22] Tha Pyai Htun, Yinxiaohe Sun, Hui Lan Chua, and Junxiong Pang. Clinical features for diagnosis of pneumonia among adults in primary care setting: A systematic and meta-review. *Scientific Reports*, 9(1):1–10, 2019.

[23] Sherif Assaad, Wolf B. Kratzert, Benjamin Shelley, Malcolm B. Friedman, and Albert Perrino. Assessment of Pulmonary Edema: Principles and Practice. *Journal of Cardiothoracic and Vascular Anesthesia*, 32(2):901–914, 2018.

[24] Akira Saito, Yukichika Hakamata, Yukiko Yamada, Mitsuhiro Sunohara, Megumi Tarui, Yoko Murano, Akihisa Mitani, Kimie Tanaka, Takahide Nagase, and Shintaro Yanagimoto. Pleural thickening on screening chest X-rays: A single institutional study. *Respiratory Research*, 20(1):1–7, 2019.

[25] Vinaya S. Karkhanis and Jyotsna M. Joshi. Pleural effusion: Diagnosis, treatment, and management. *Open Access Emergency Medicine*, 4:31–52, 2012.

[26] Jong Jin Hyun and Young Tae Bak. Clinical significance of hiatal hernia, sep 2011.

[27] I. Petrache and K. Serban. Emphysema. In *Pathobiology of Human Disease: A Dynamic Encyclopedia of Disease Mechanisms*, pages 2609–2624. Elsevier Inc., jan 2014.

[28] N. J. Shaw, M. Hendry, and O. B. Eden. Inter-observer variation in interpretation of chest X-rays. *Scottish Medical Journal*, 35(5):140–141, 1990.

[29] Koichiro Yasaka and Osamu Abe. Deep learning and artificial intelligence in radiology: Current applications and future directions. *PLOS Medicine*, 15(11):e1002707, nov 2018.

[30] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-January:3462–3471, may 2017.

[31] Bram van Ginneken. Fifty years of computer analysis in chest imaging: rule-based, machine learning, deep learning. *Radiological Physics and Technology*, 10(1):23–32, 2017.

[32] Expert.ai Team. What is Machine Learning? A definition - Expert System | Expert.ai, 2017.

[33] Chunli Qin, Demin Yao, Yonghong Shi, and Zhijian Song. Computer-aided detection in chest radiography based on artificial intelligence: A survey, aug 2018.

[34] Lea Pehrson, Michael Nielsen, and Carsten Ammitzbøl Lauridsen. Automatic Pulmonary Nodule Detection Applying Deep Learning or Machine Learning Algorithms to the LIDC-IDRI Database: A Systematic Review. *Diagnostics*, 9(1):29, mar 2019.

[35] Ali H. Al-Timemy, Rami N. Khushaba, Zahraa M. Mosa, and Javier Escudero. An Efficient Mixture of Deep and Machine Learning Models for COVID-19 and Tuberculosis Detection Using X-Ray Images in Resource Limited Settings. *arXiv*, jul 2020.

[36] François Chollet. *Deep Learning with Python*. Manning Publications Co., 2018.

[37] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Institute of Electrical and Electronics Engineers (IEEE), 2010.

[38] E. J. Yates, L. C. Yates, and H. Harvey. Machine learning "red dot": open-source, cloud, deep convolutional neural networks in chest radiograph binary normality classification. *Clinical Radiology*, 73(9):827–831, sep 2018.

[39] Qingji Guan, Yaping Huang, Zhun Zhong, Zhedong Zheng, Liang Zheng, and Yi Yang. Diagnose like a Radiologist: Attention Guided Convolutional Neural Network for Thorax Disease Classification. *arXiv*, jan 2018.

[40] Bingzhi Chen, Jinxing Li, Xiaobao Guo, and Guangming Lu. DualCheXNet: dual asymmetric feature learning for thoracic disease classification in chest X-rays. *Biomedical Signal Processing and Control*, 53:101554, aug 2019.

[41] Cheng Zhang, Francine Chen, and Yan-Ying Chen. Thoracic Disease Identification and Localization using Distance Learning and Region Verification. *arXiv*, jun 2020.

[42] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, pages 590–597, jan 2019.

[43] Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv*, jan 2019.

[44] Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu H. Pham, Dat Q. Tran, Dung B. Nguyen, Dung D. Le, Chi M. Pham, Hang T. T. Tong, Diep H. Dinh, Cuong D. Do, Luu T. Doan, Cuong N. Nguyen, Binh T. Nguyen, Que V. Nguyen, Au D. Hoang, Hien N. Phan, Anh T. Nguyen, Phuong H. Ho, Dat T. Ngo, Nghia T. Nguyen, Nhan T. Nguyen, Minh Dao, and Van Vu. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. 2020.

[45] Chaochao Yan, Jiawen Yao, Ruoyu Li, Zheng Xu, and Junzhou Huang. Weakly Supervised Deep Learning for Thoracic Disease Classification and Localization on Chest X-rays. *ACM-BCB 2018 - Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 18:103–110, jul 2018.

[46] Alistair E.W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):1–8, dec 2019.

[47] Luke Oakden-Rayner. Exploring the ChestXray14 dataset: problems – Luke Oakden-Rayner, 2017.

[48] Hoo Chang Shin, Le Lu, and Ronald M. Summers. Natural Language Processing for Large-Scale Medical Image Analysis Using Deep Learning. In *Deep Learning for Medical Image Analysis*, pages 405–421. Elsevier Inc., jan 2017.

[49] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.

[50] Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan. Explainable deep learning models in medical image analysis. Technical report, Theoretical and Experimental Epistemology Laboratory, School of Optometry and Vision Science, University of Waterloo, Ontario, Canada; Department of Systems Design Engineering, University of Waterloo, Ontario, Canada, 2020.

[51] Holger R. Roth, Le Lu, Jiamin Liu, Jianhua Yao, Ari Seff, Kevin Cherry, Lauren Kim, and Ronald M. Summers. Improving Computer-Aided Detection Using Convolutional Neural Networks and Random View Aggregation. *IEEE Transactions on Medical Imaging*, 35(5):1170–1181, 2016.

[52] Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, 58, dec 2019.

[53] Diederik P Kingma Google, Max Welling, and Boston Delft. An Introduction to Variational Autoencoders. *Foundations and Trends R in Machine Learning*, xx, No. xx:1–18, 2019.

[54] Rowel Atienza. *Advanced Deep Learning with TensorFlow 2 and Keras_ Apply DL, GANs, VAEs, deep RL, unsupervised learning, object detection and segmentation, and more, 2nd Edition-Packt Publishing*. Number 2. Packt Publishing, 2019.

[55] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. *arXiv*, 2014.

[56] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, pages 1–16, 2016.

[57] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, dec 2015.

[58] Yann LeCun, Corinna Cortes, and Chris Burges. MNIST handwritten digit database.

[59] F Munawar, S Azmat, T Iqbal, C Grönlund, and H Ali. Segmentation of Lungs in Chest X-Ray Image Using Generative Adversarial Networks. *IEEE Access*, 2020.

[60] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional Image Synthesis With Auxiliary Classifier GANs. *34th International Conference on Machine Learning, ICML 2017*, 6:4043–4055, oct 2016.

[61] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. *Advances in Neural Information Processing Systems*, pages 2180–2188, jun 2016.

[62] Tero Karras, Weili Nie, Animesh Garg, Shoubhik Debnath, Anjul Patney, Ankit B. Patel, and Anima Anandkumar. Semi-supervised styleGAN for disentanglement learning. *arXiv*, 2020.

[63] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation, oct 2017.

[64] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A Neural Algorithm of Artistic Style. *Journal of Vision*, 16(12):326, aug 2015.

[65] Xun Huang and Serge Belongie. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-October:1510–1519, mar 2017.

[66] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and Improving the Image Quality of StyleGAN. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 8107–8116, dec 2019.

[67] Bingchen Liu, Yizhe Zhu, Kunpeng Song, and Ahmed Elgammal. TOWARDS FASTER AND STABILIZED GAN TRAINING FOR HIGH-FIDELITY FEW-SHOT IMAGE SYNTHESIS. 2021.

[68] Esther Robb, Wen-Sheng Chu, Abhishek Kumar, and Jia-Bin Huang. Few-Shot Adaptation of Generative Adversarial Networks. oct 2020.

[69] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local Neural Networks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 7794–7803. IEEE Computer Society, dec 2018.

[70] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze the Discriminator: a Simple Baseline for Fine-Tuning GANs. feb 2020.

[71] Atsuhiro Noguchi and Tatsuya Harada. Image Generation From Small Datasets via Batch Statistics Adaptation. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-October:2750–2758, apr 2019.

[72] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. SinGAN: Learning a generative model from a single natural image. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2019-October, pages 4569–4579. Institute of Electrical and Electronics Engineers Inc., oct 2019.

[73] Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, Brian Curless, and David H Salesin. Image Analogies. Technical report, New York University, Microsoft Research, University of Washington, 2001.

[74] Pramuditha Perera, Mahdi Abavisani, and Vishal M. Patel. In2I : Unsupervised Multi-Image-to-Image Translation Using Generative Adversarial Networks. *Proceedings - International Conference on Pattern Recognition*, 2018-August:140–146, nov 2017.

[75] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9351, pages 234–241. Springer Verlag, may 2015.

[76] Phillip Isola, Jun Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:5967–5976, 2017.

[77] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved Techniques for Training GANs. *30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain*, 2016.

[78] David Warde-Farley and Ian Goodfellow. Adversarial Perturbations of Deep Neural Networks. In *Perturbations, Optimization, and Statistics*. The MIT Press, dec 2018.

[79] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *arXiv*, jan 2017.

[80] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang, and Stephen Paul Smolley. Least Squares Generative Adversarial Networks. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-October:2813–2821, nov 2016.

[81] Ali Borji. Pros and cons of GAN evaluation measures. *Computer Vision and Image Understanding*, 179:41–65, feb 2019.

[82] Donald Geman, Stuart Geman, Neil Hallonquist, and Laurent Younes. Visual Turing test for computer vision systems. *Proceedings of the National Academy of Sciences of the United States of America*, 112(12):3618–3623, mar 2015.

[83] Maria J.M. Chuquicusma, Sarfaraz Hussein, Jeremy Burt, and Ulas Bagci. How to fool radiologists with generative adversarial networks? A visual turing test for lung cancer diagnosis. In *Proceedings - International Symposium on Biomedical Imaging*, volume 2018-April, pages 240–244. IEEE Computer Society, may 2018.

[84] Shane Barratt and Rishi Sharma. A Note on the Inception Score. *arXiv*, jan 2018.

[85] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *Advances in Neural Information Processing Systems*, 2017-December:6627–6638, jun 2017.

[86] Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, jan 2018.

[87] Neeraj Kumar, Srishti Goel, Ankur Narang, Brejesh Lall, Mujtaba Hasan, Pranshu Agarwal, and Dipankar Sarkar. One Shot Audio to Animated Video Generation. feb 2021.

[88] Youssef A. Mejjati, Christian Richardt, James Tompkin, Darren Cosker, and Kwang In Kim. Unsupervised Attention-guided Image to Image Translation. *Advances in Neural Information Processing Systems*, 2018-December:3693–3703, jun 2018.

[89] Ramyasree Kola. *Generation of synthetic plant images using deep learning architecture*. PhD thesis, Blekinge Institute of Technology, 2019.

[90] Shuyue Guan and Murray Loew. A Novel Measure to Evaluate Generative Adversarial Networks Based on Direct Analysis of Generated Images. *arXiv*, feb 2020.

[91] David Lopez-Paz and Maxime Oquab. Revisiting Classifier Two-Sample Tests. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, oct 2016.

[92] Qiantong Xu, Gao Huang, Yang Yuan, Chuan Guo, Yu Sun, Felix Wu, and Kilian Weinberger. An empirical study on evaluation metrics of generative adversarial networks. *arXiv*, jun 2018.

[93] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *arXiv*, pages 4401–4410, 2018.

[94] Benyamin Ghojogh, Fakhri Karray, and Mark Crowley. Theoretical Insights into the Use of Structural Similarity Index In Generative Models and Inferential Autoencoders. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12132 LNCS:112–117, apr 2020.

[95] Tong Che, Yanran Li, Athul Paul Jacob, Yoshua Bengio, and Wenjie Li. Mode Regularized Generative Adversarial Networks. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, dec 2016.

[96] Vera Sorin, Yiftach Barash, Eli Konen, and Eyal Klang. Creating Artificial Images for Radiology Applications Using Generative Adversarial Networks (GANs) – A Systematic Review. *Academic Radiology*, 27(8):1175–1185, 2020.

[97] Web of Science [v.5.35] - Principal Coleção do Web of ScienceResultados.

[98] Sagar Kora Venu and Sridhar Ravula. Evaluation of deep convolutional generative adversarial networks for data augmentation of chest x-ray images. *Future Internet*, 13(1):1–13, 2021.

[99] Abdul Waheed, Muskan Goyal, Deepak Gupta, Ashish Khanna, Fadi Al-Turjman, and Placido Rogerio Pinheiro. CovidGAN: Data Augmentation Using Auxiliary Classifier GAN for Improved Covid-19 Detection. *IEEE Access*, 8:91916–91923, 2020.

[100] Mehdi Moradi, Ali Madani, Alexandros Karargyris, and Tanveer F. Syeda-Mahmood. Chest x-ray generation and data augmentation for cardiovascular abnormality classification. In *SPIE Medical Imaging*, number 10574, page 57. SPIE, 2018.

[101] Jia Liang, Yu-Xing Tang, You-Bao Tang, Jing Xiao, and Ronald M Summers. Bone Suppression on Chest Radiographs With Adversarial Learning. *arXiv*, page 6, 2020.

[102] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, sep 2015.

[103] Alex J. DeGrave, Joseph D. Janizek, and Su In Lee. AI for radiographic COVID-19 detection selects shortcuts over signal, oct 2020.

[104] VinDr-CXR: An open dataset and benchmarks for disease classification and abnormality localization on chest radiographs | VinDr.

[105] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[106] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-October:5908–5916, 2017.

[107] Animesh Karnewar and Oliver Wang. MSG-GAN: Multi-Scale Gradients for Generative Adversarial Networks. *arXiv*, 2020.

[108] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 07-12-June-2015, pages 1–9. IEEE Computer Society, oct 2015.

[109] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *arXiv*, 2015.

[110] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-Excitation Networks. *arXiv*, 2019.

[111] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance Normalization: The Missing Ingredient for Fast Stylization. *arXiv*, jul 2016.

[112] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and Benchmarking Self-Supervised Visual Representation Learning. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-October:6390–6399, may 2019.

[113] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 9726–9735, nov 2019.

[114] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty. *Advances in Neural Information Processing Systems*, 32, jun 2019.

[115] Longlong Jing and Yingli Tian. Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, feb 2019.

[116] Zhuoran Shen, Irwan Bello, Raviteja Vemulapalli, Xuhui Jia, and Ching-Hui Chen. Global Self-Attention Networks For Image Recorgnition. *ICLR 2021 Conference Blind Submission*, sep 2020.

[117] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V. Le. Attention Augmented Convolutional Networks. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-October:3285–3294, apr 2019.

[118] Kaiyu Yue, Ming Sun, Yuchen Yuan, Feng Zhou, Errui Ding, and Fuxin Xu. Compact Generalized Non-local Network. *Advances in Neural Information Processing Systems*, 2018-December:6510–6519, oct 2018.

[119] Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient Attention: Attention with Linear Complexities. dec 2018.

[120] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. A 2-Nets: Double Attention Networks. Technical report.

[121] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *32nd International Conference on Machine Learning, ICML 2015*, volume 1, pages 448–456. International Machine Learning Society (IMLS), feb 2015.

[122] Jae Hyun Lim and Jong Chul Ye. Geometric GAN. may 2017.

[123] Dustin Tran, Rajesh Ranganath, and David M. Blei. Hierarchical Implicit Models and Likelihood-Free Variational Inference. *Advances in Neural Information Processing Systems*, 2017-December:5524–5534, feb 2017.

[124] Ning Yu, Guilin Liu, Aysegul Dundar, Andrew Tao, Bryan Catanzaro, Larry Davis, and Mario Fritz. Dual Contrastive Loss and Attention for GANs. mar 2021.

[125] Minguk Kang and Jaesik Park. ContraGAN: Contrastive Learning for Conditional Image Generation. Technical report, jun 2020.

[126] Zhengli Zhao, Zizhao Zhang, Ting Chen, Sameer Singh, and Han Zhang. Image Augmentations for GAN Training. jun 2020.

[127] Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive Learning for Unpaired Image-to-Image Translation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12354 LNCS:319–345, jul 2020.

[128] Carl F. Sabottke and Bradley M. Spieler. The Effect of Image Resolution on Deep Learning in Radiography. *Radiology: Artificial Intelligence*, 2(1):e190015, jan 2020.

[129] Yasin Yazıcı, Chuan-Sheng Foo, Stefan Winkler, Kim-Hui Yap, Georgios Piliouras, and Vijay Chandrasekhar. The Unusual Effectiveness of Averaging in GAN Training. *7th International Conference on Learning Representations, ICLR 2019*, jun 2018.

[130] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. apr 2017.

[131] Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, Ayush Chaurasia, TaoXie, Liu Changyu, Abhiram V, Laughing, Tkianai, YxNONG, Adam Hogan, Lorenzomammana, AlexWang1900, Jan Hajek, Laurentiu Diaconu, Marc, Yonghye Kwon, Oleg, Wanghaoyang0106, Yann Defretin, Aditya Lohia, Ml5ah, Ben Milanko, Benjamin Fineran, Daniel Khromov, Ding Yiwei, Doug, Durgesh, and Francisco Ingham. ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations. apr 2021.

Appendix A

Additional Results

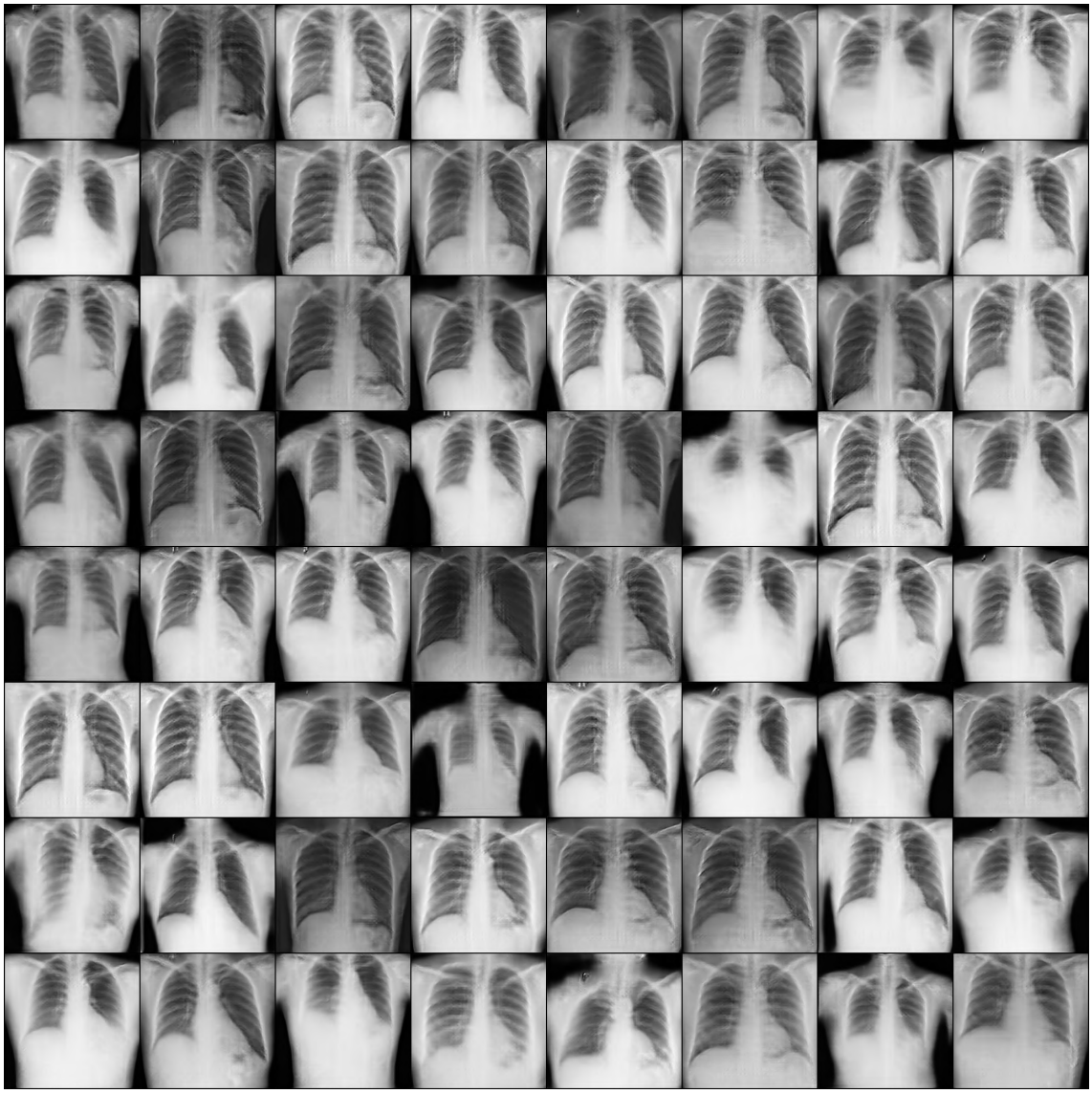


Figure A.1: 64 generated samples by the *256 Resolution* model

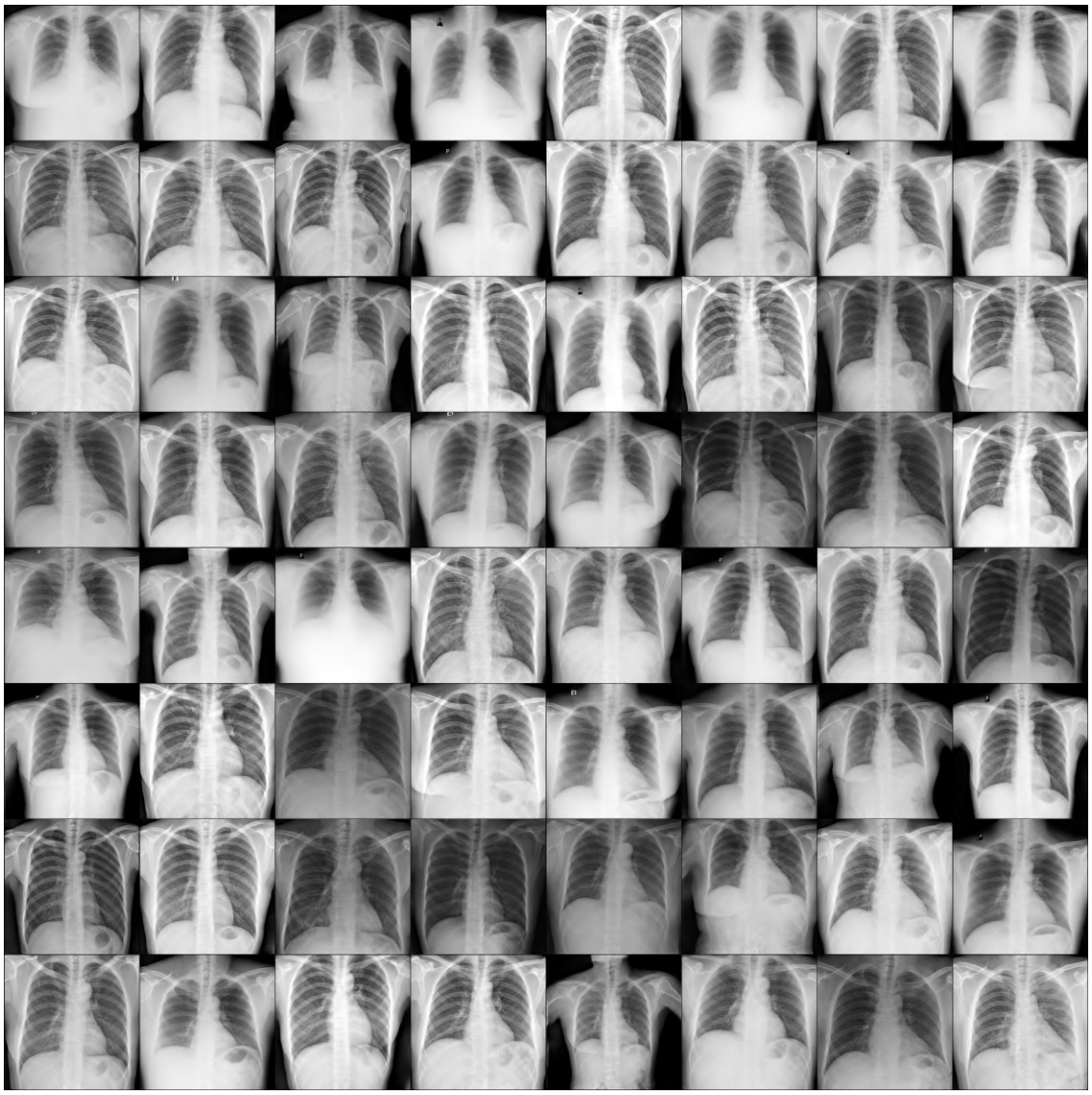


Figure A.2: 64 generated samples by the *512 Resolution* model

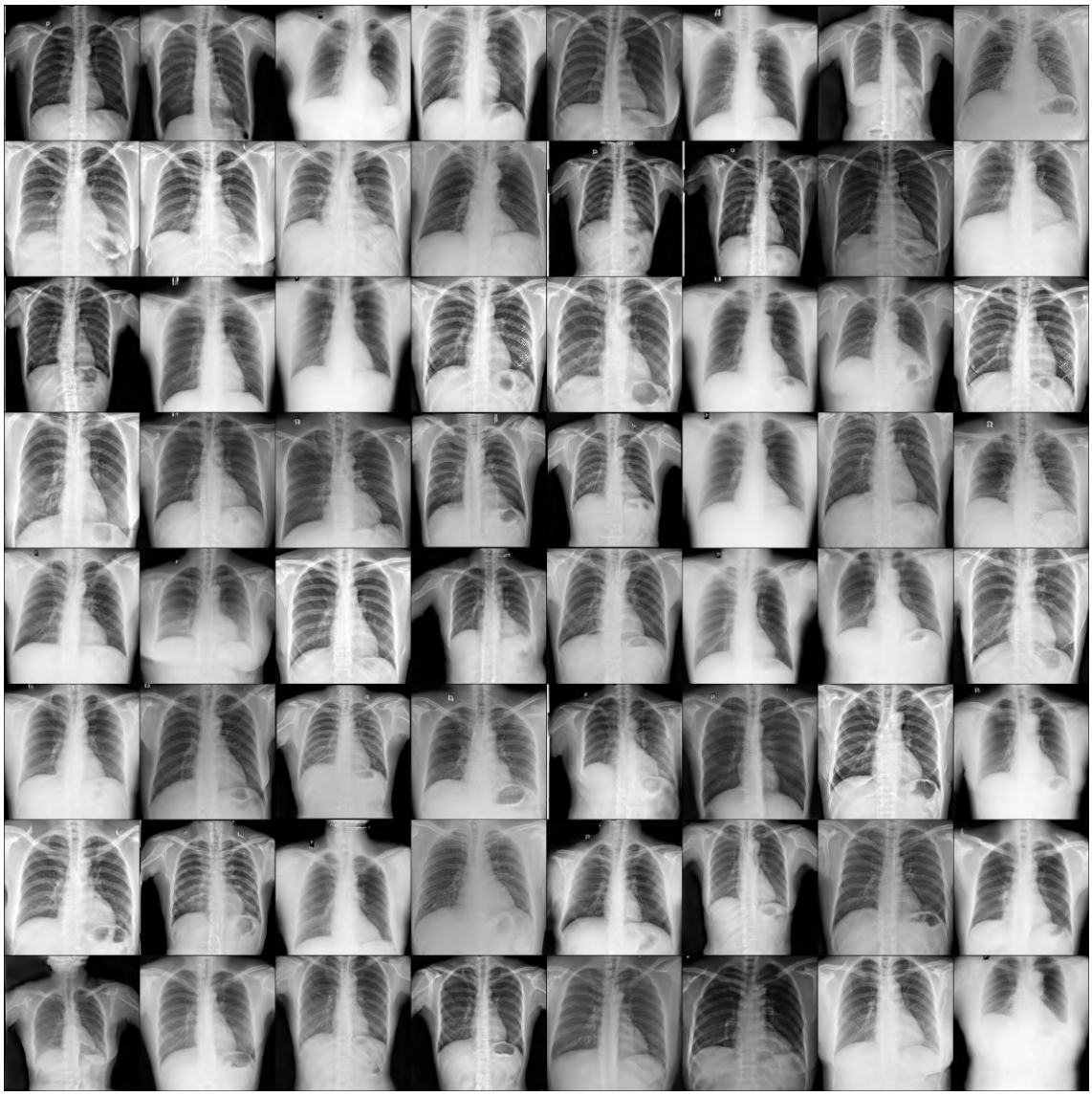


Figure A.3: 64 generated samples by the *32 GSA* model

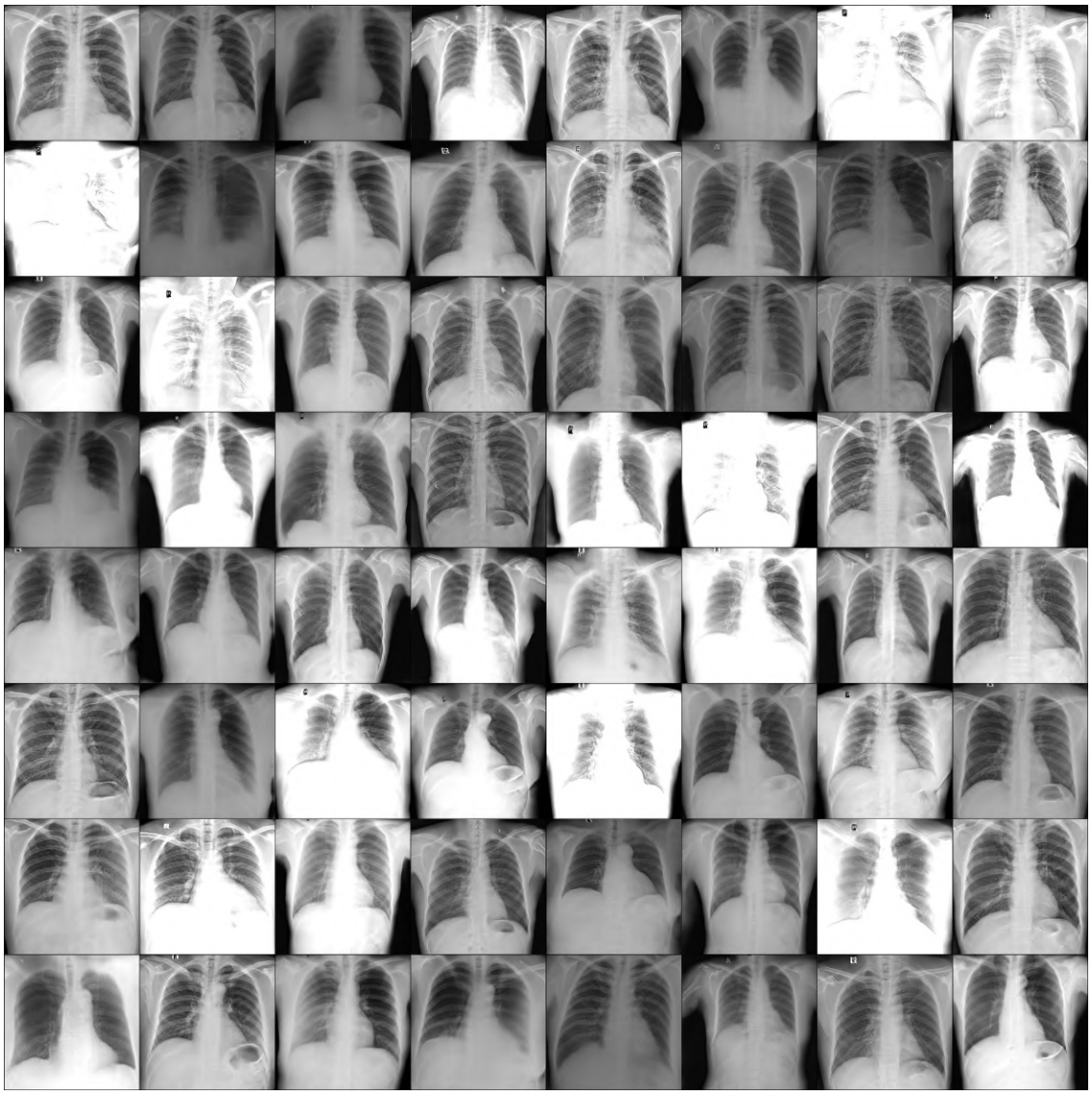


Figure A.4: 64 generated samples by the *256 GSA* model

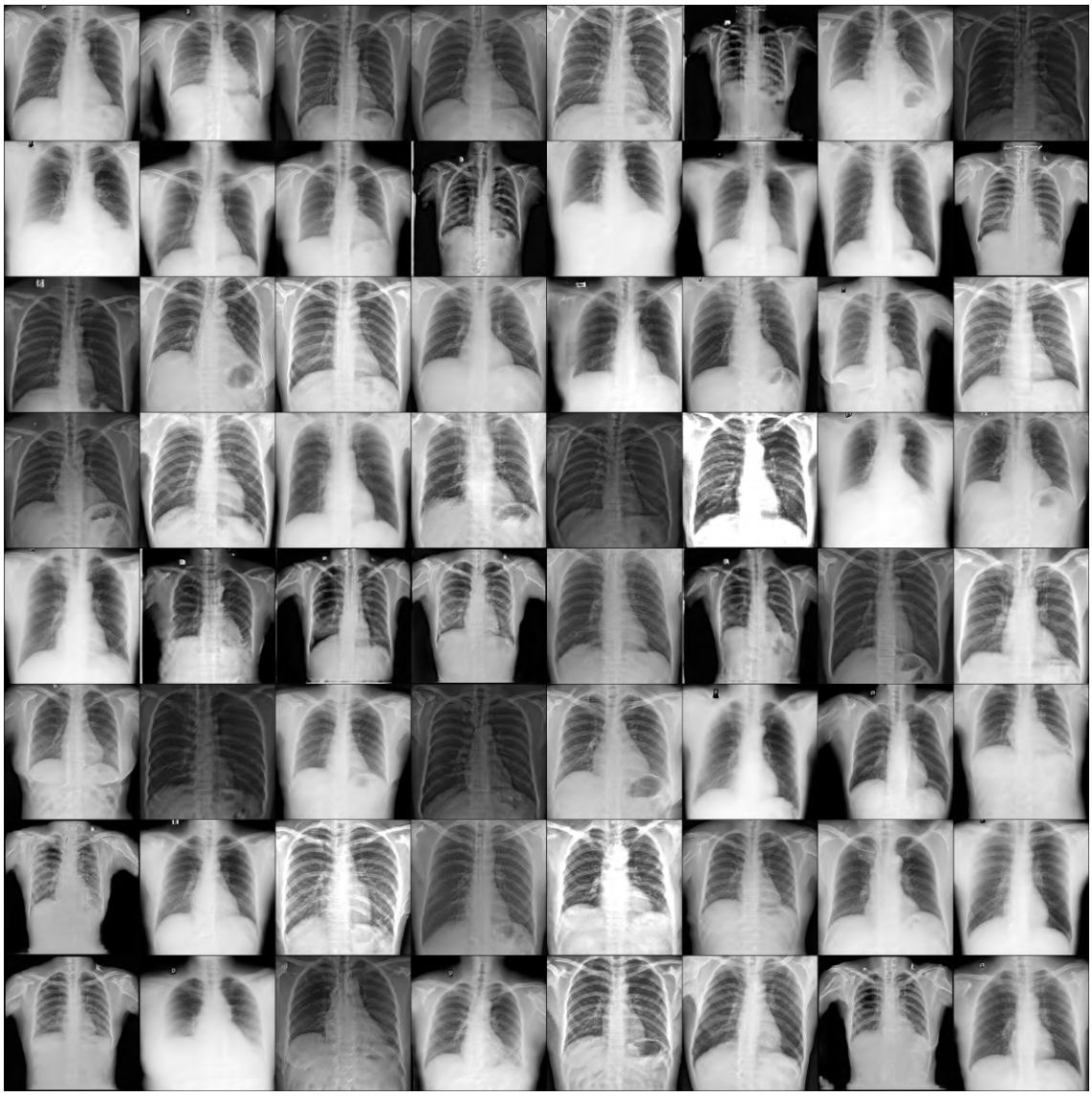


Figure A.5: 64 generated samples by the *32-128 GSA* model

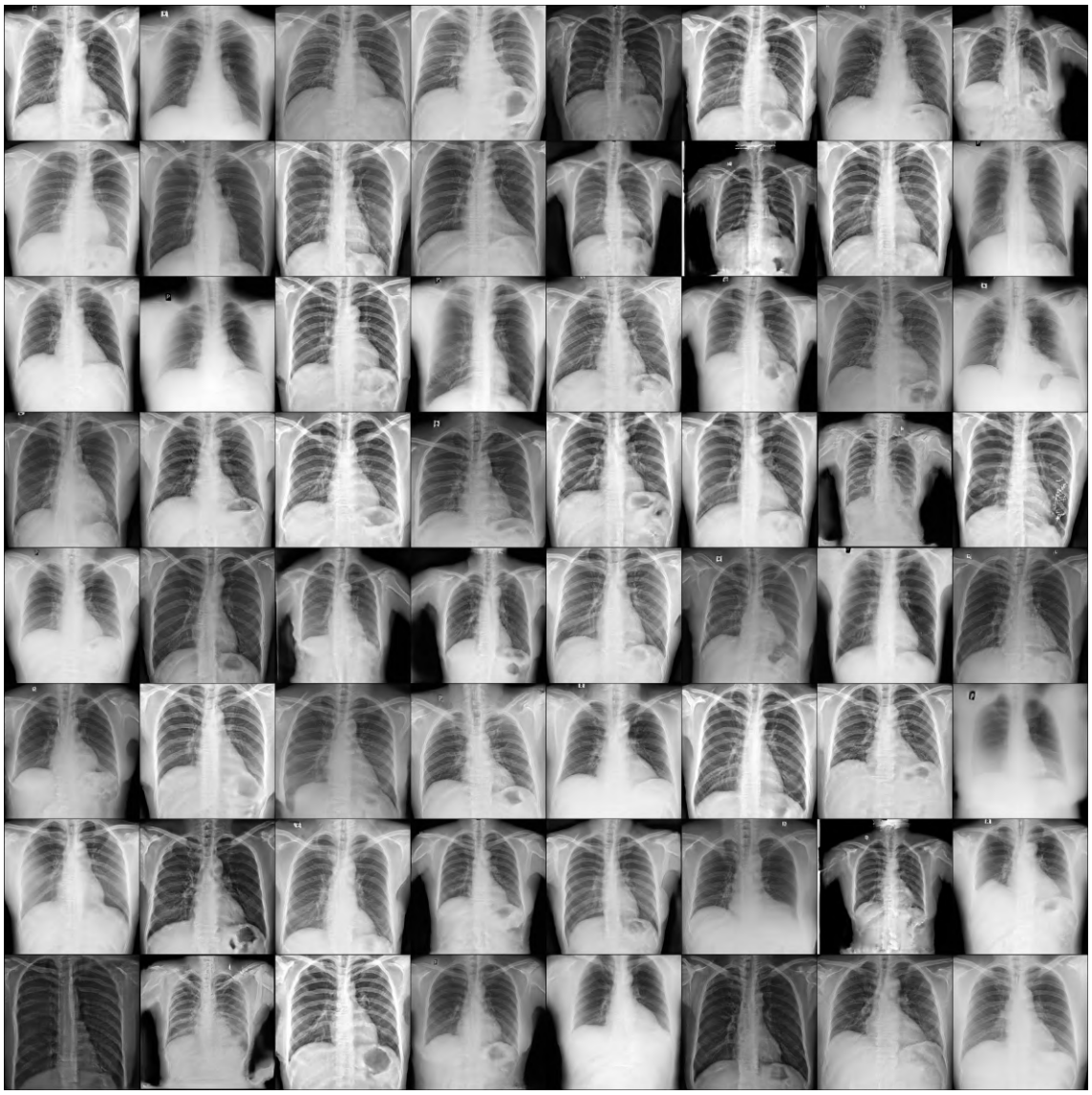


Figure A.6: 64 generated samples by the *All Layers* model

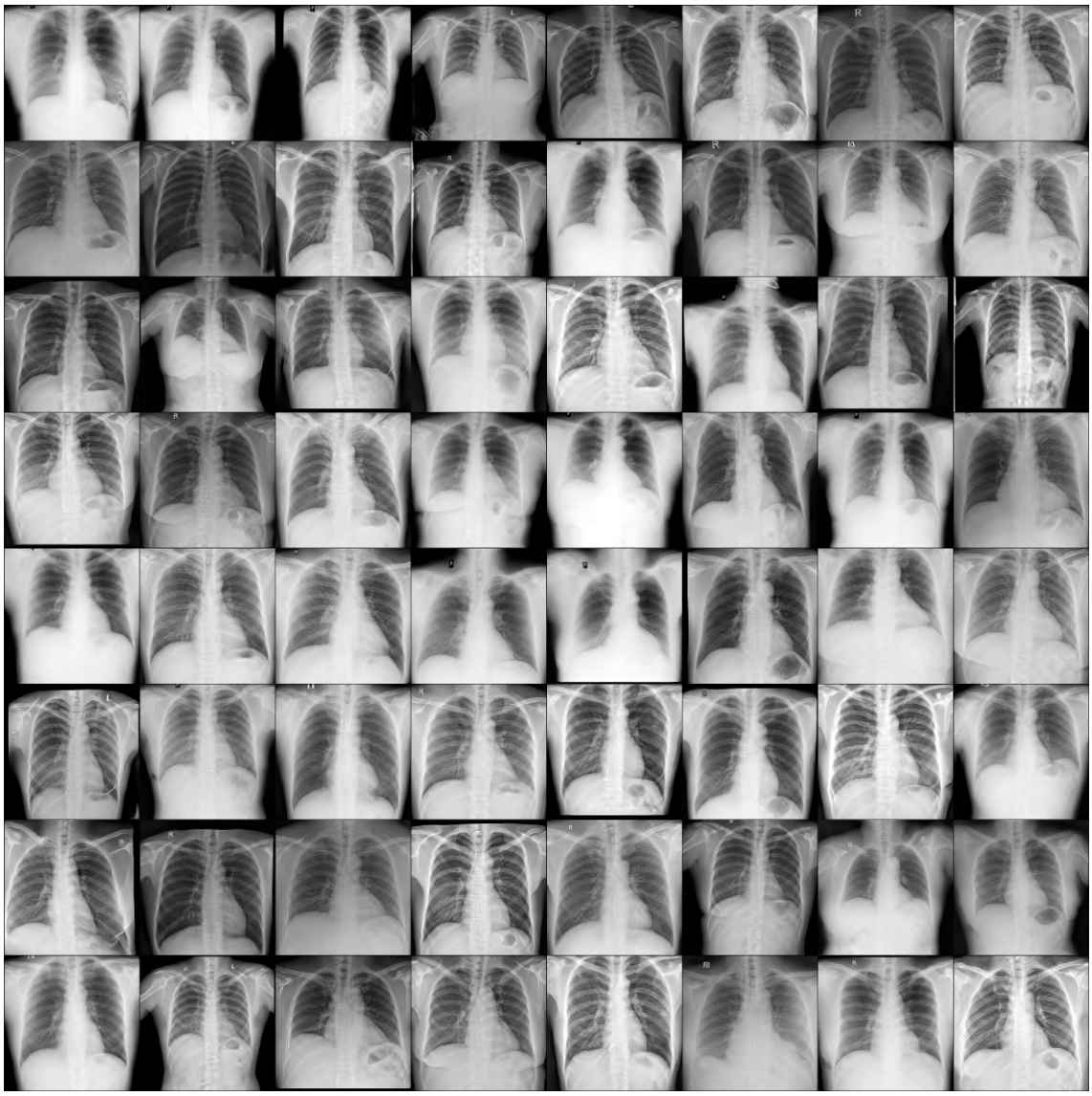


Figure A.7: 64 generated samples by the *Dual Contrastive Loss* model

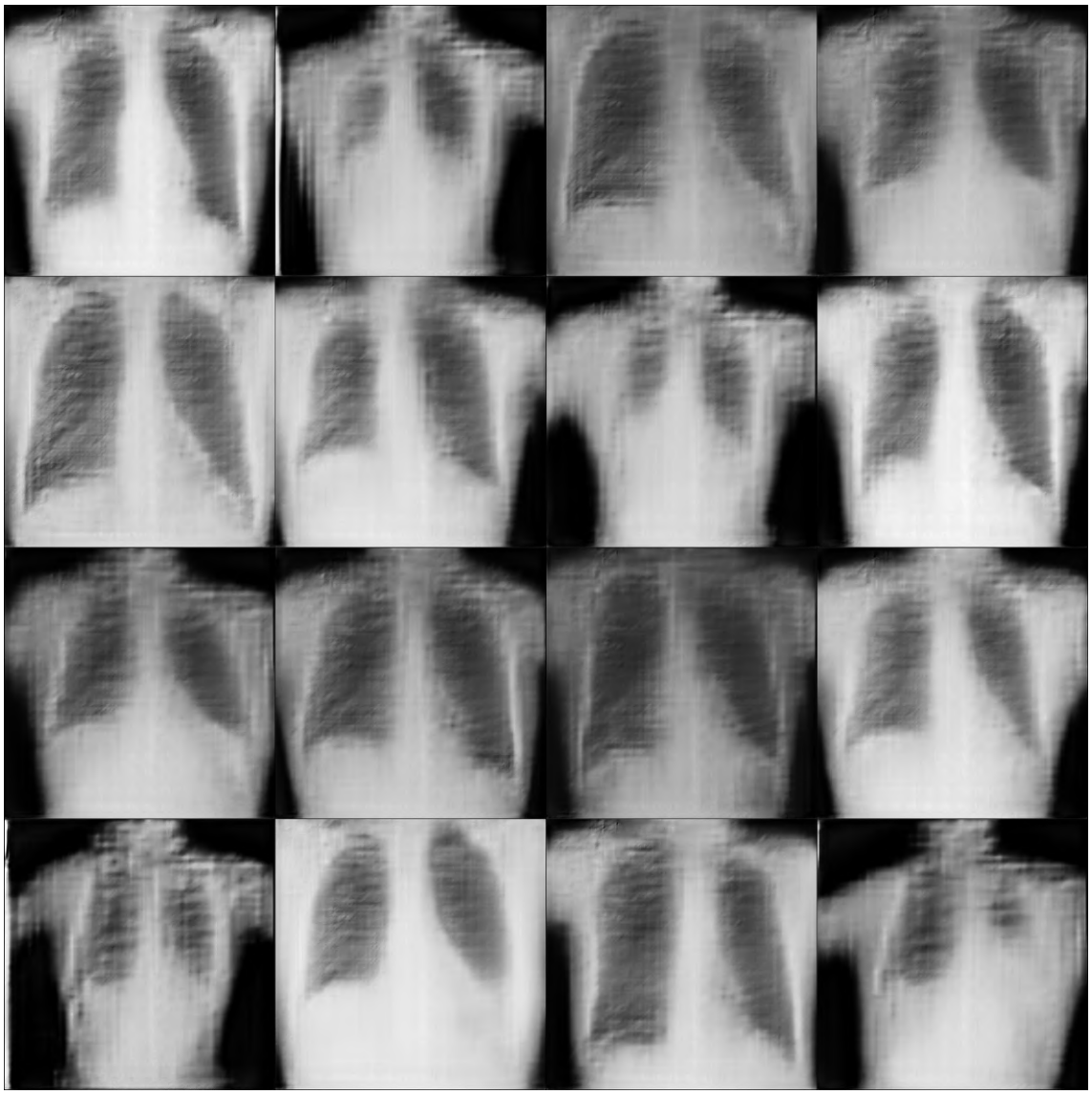


Figure A.8: 16 generated samples by the *1024 Resolution* model

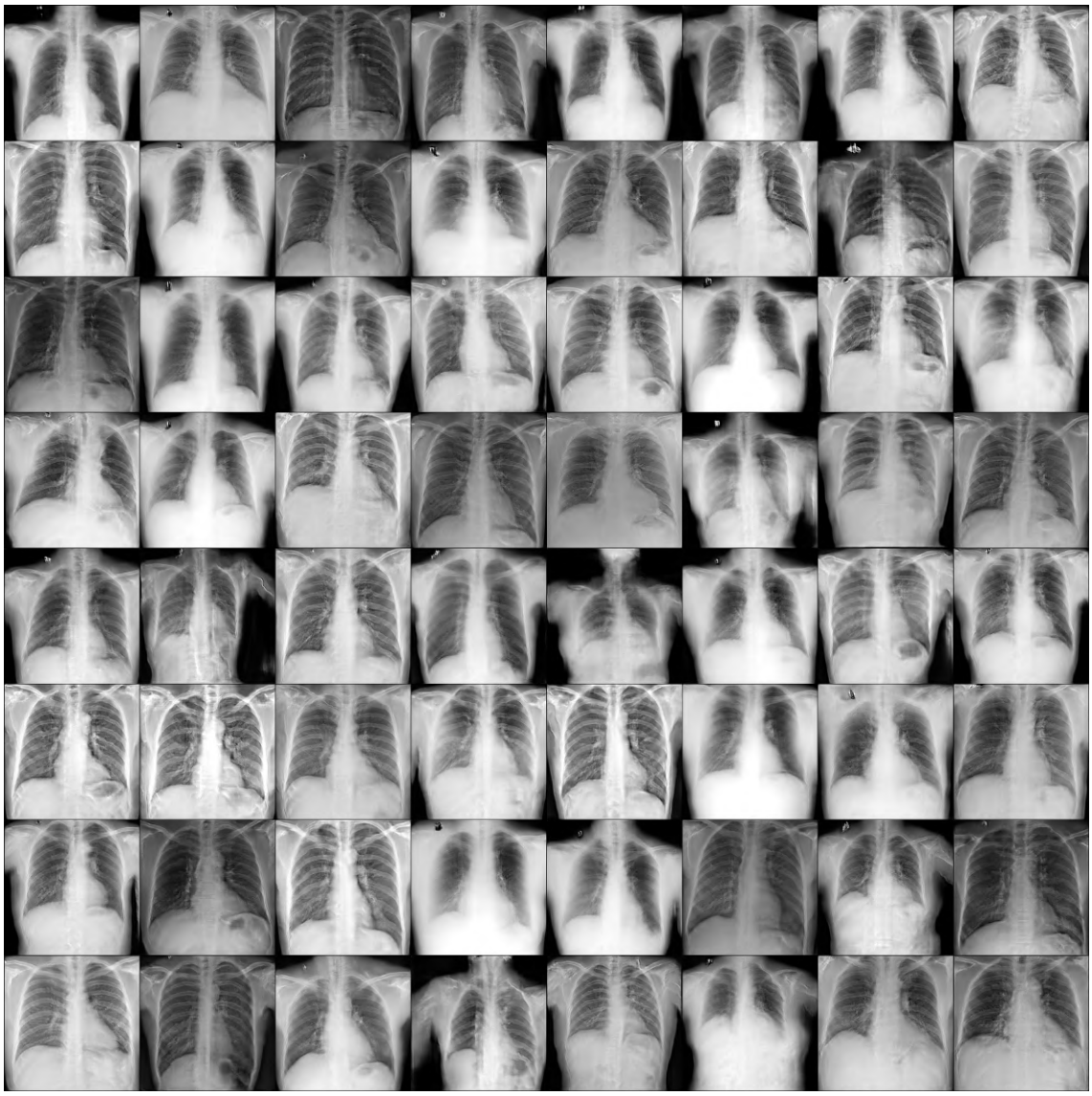


Figure A.9: 64 generated samples by the *1000 Image Set* model

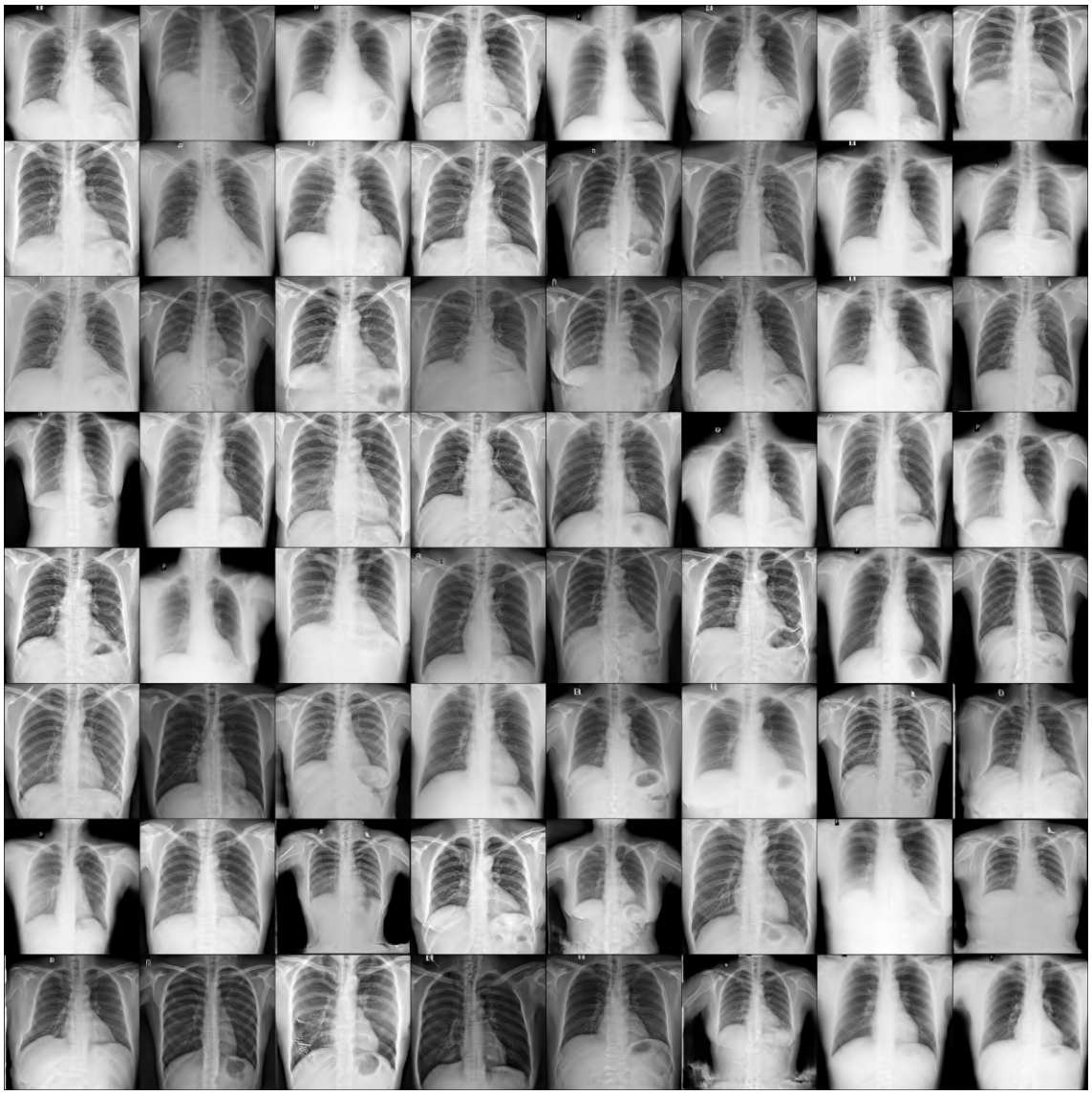


Figure A.10: 64 generated samples by the *4000 Image Set* model

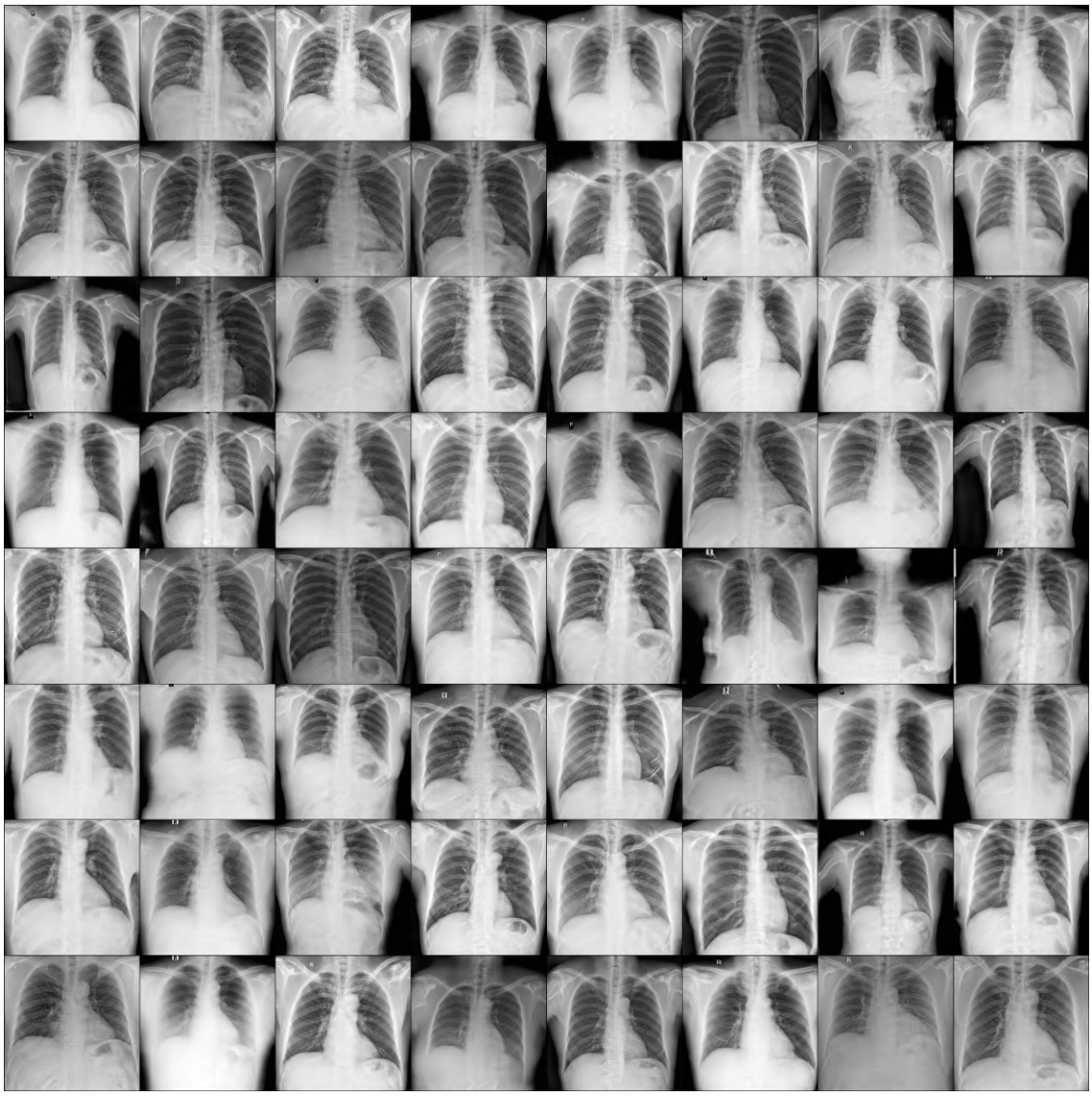


Figure A.11: 64 generated samples by the *4000 Image Set - Larger Batch Size* model

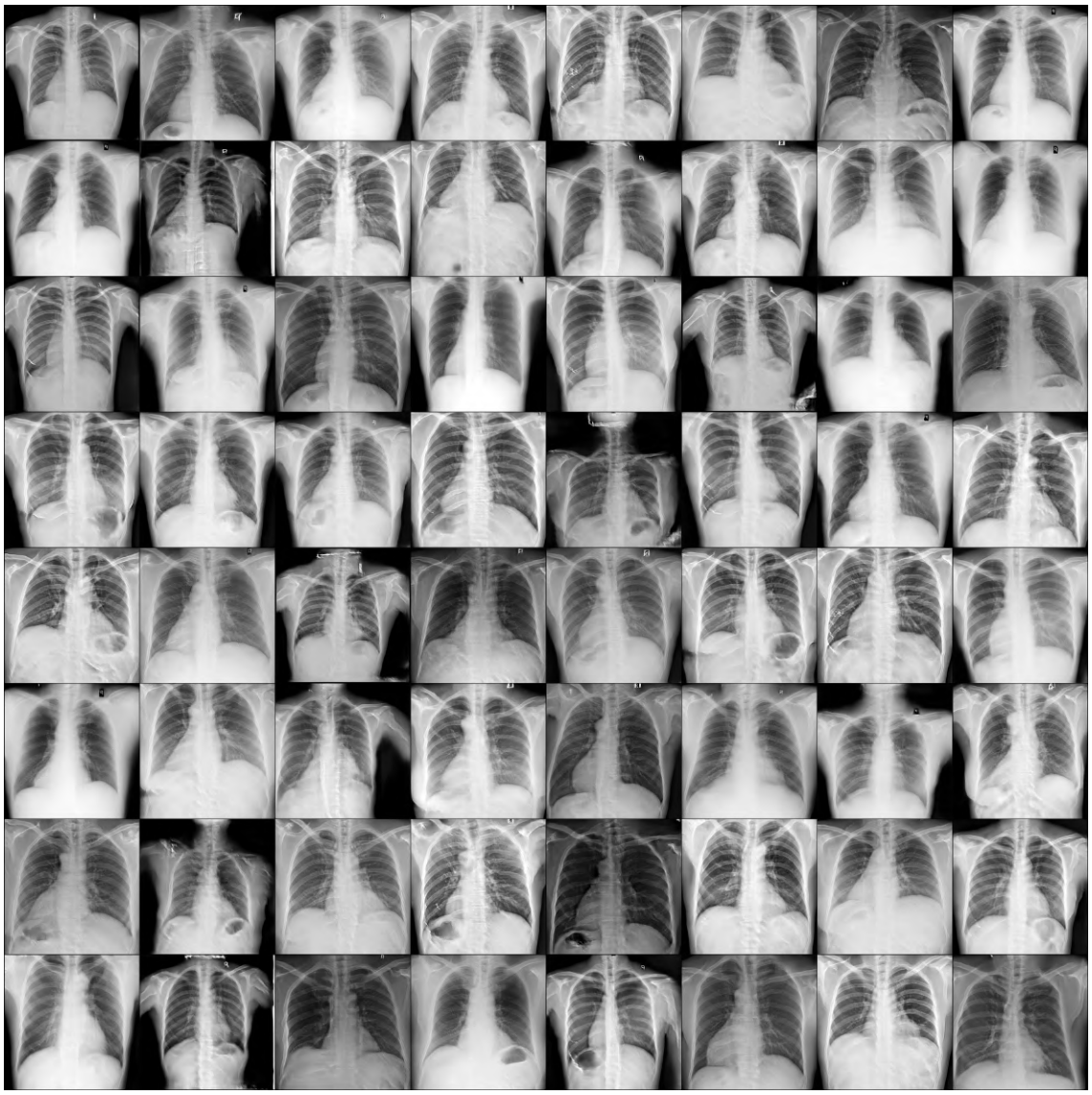


Figure A.12: 64 generated samples by the *4000 Image Set - Horizontal Flip* $p = 0.5$ model

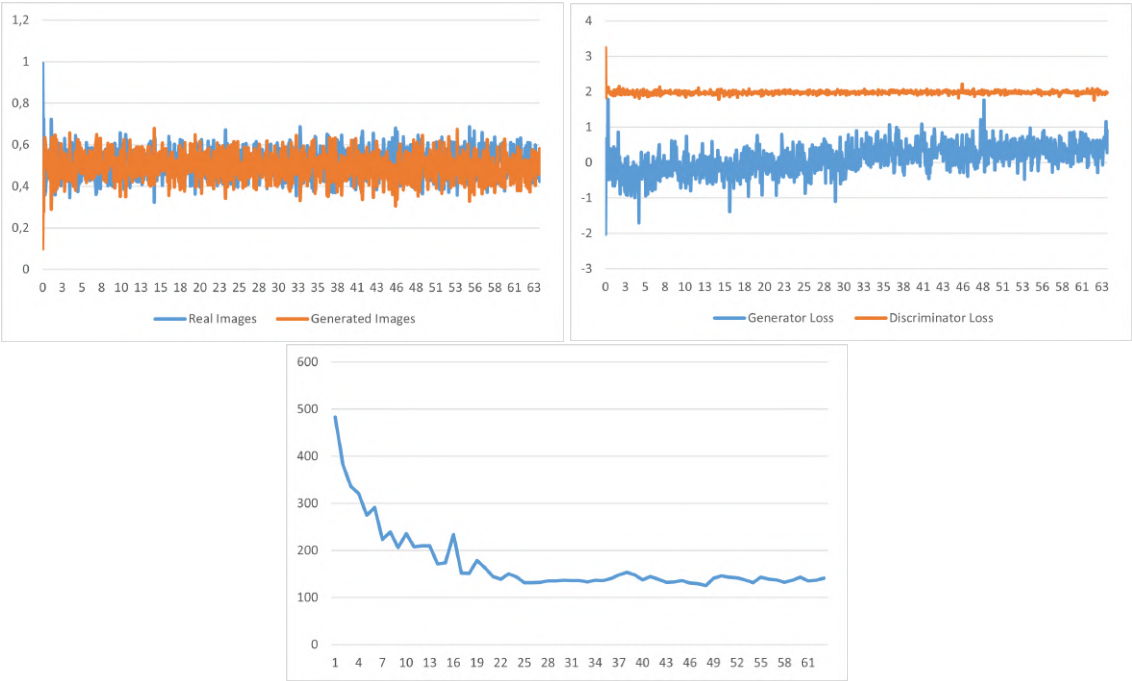


Figure A.13: Discriminator accuracy at predicting the image source (left), loss functions (right) and FID (center) values from the *256 Resolution* model throughout training

Figure A.14: Discriminator accuracy at predicting the image source (left), loss functions (right) and FID (center) values from the *512 Resolution* model throughout training

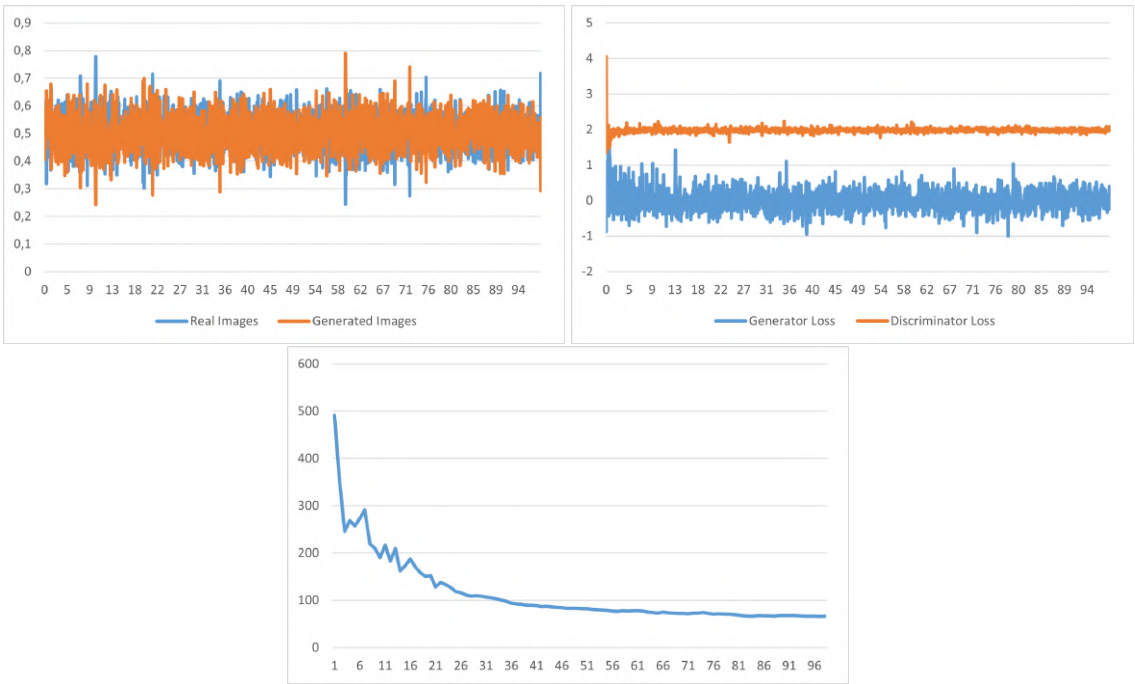


Figure A.15: Discriminator accuracy at predicting the image source (left), loss functions (right) and FID (center) values from the *256 GSA* model throughout training

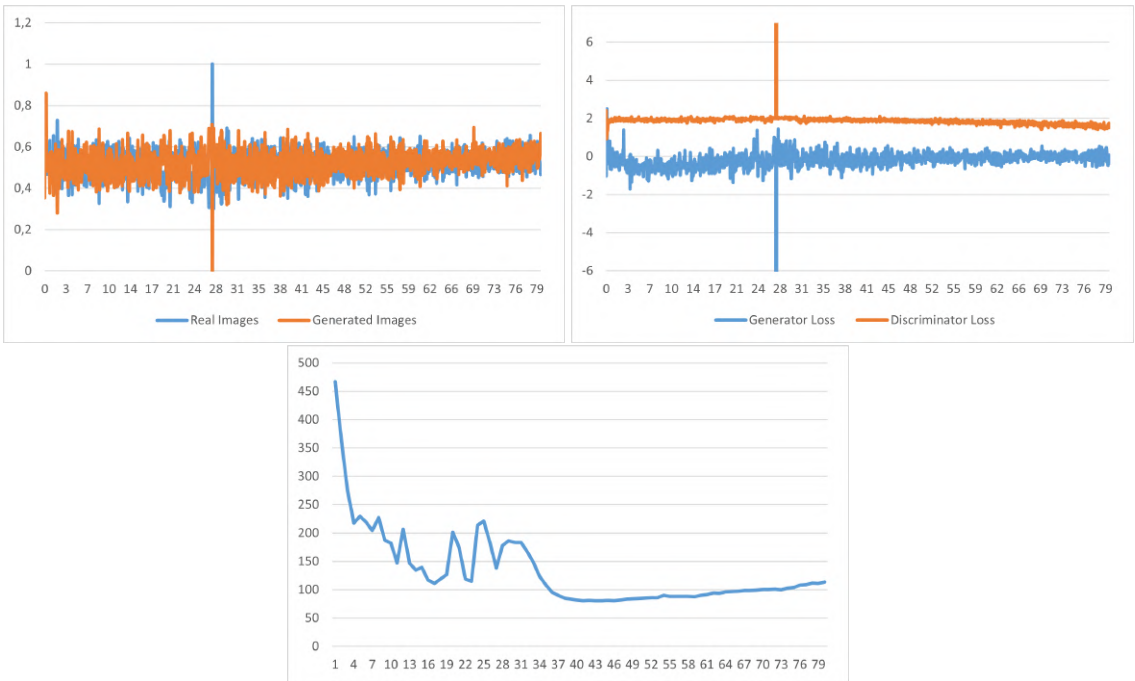


Figure A.16: Discriminator accuracy at predicting the image source (left), loss functions (right) and FID (center) values from the *32-128 GSA* model throughout training

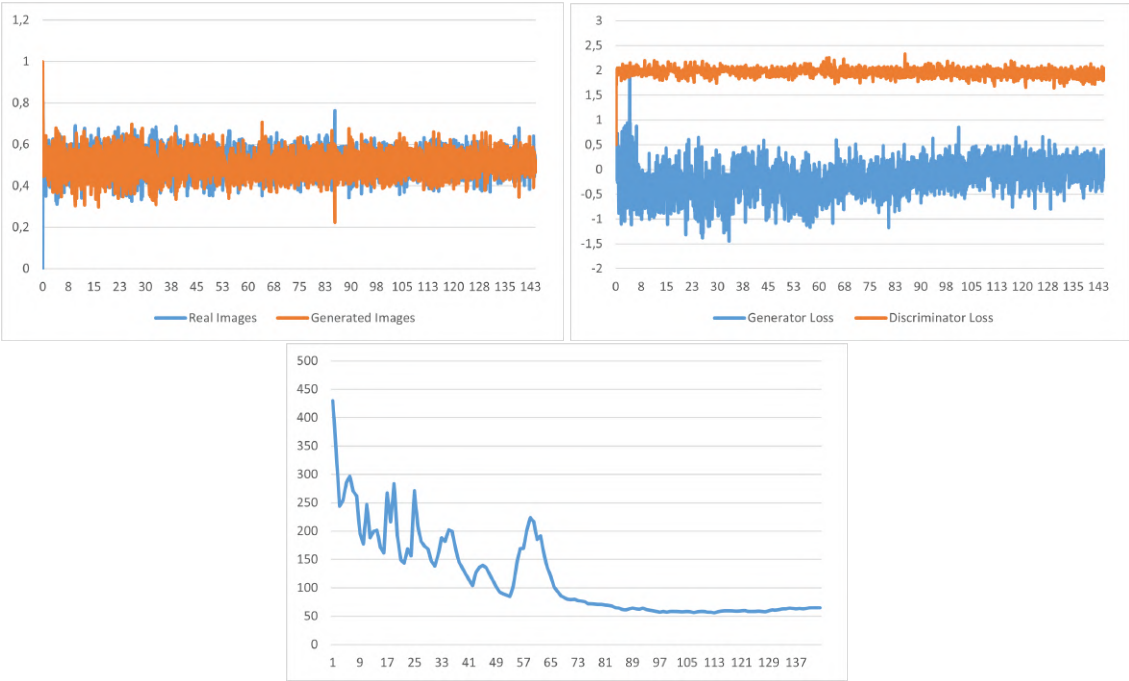


Figure A.17: Discriminator accuracy at predicting the image source (left), loss functions (right) and FID (center) values from the *All Layers* model throughout training

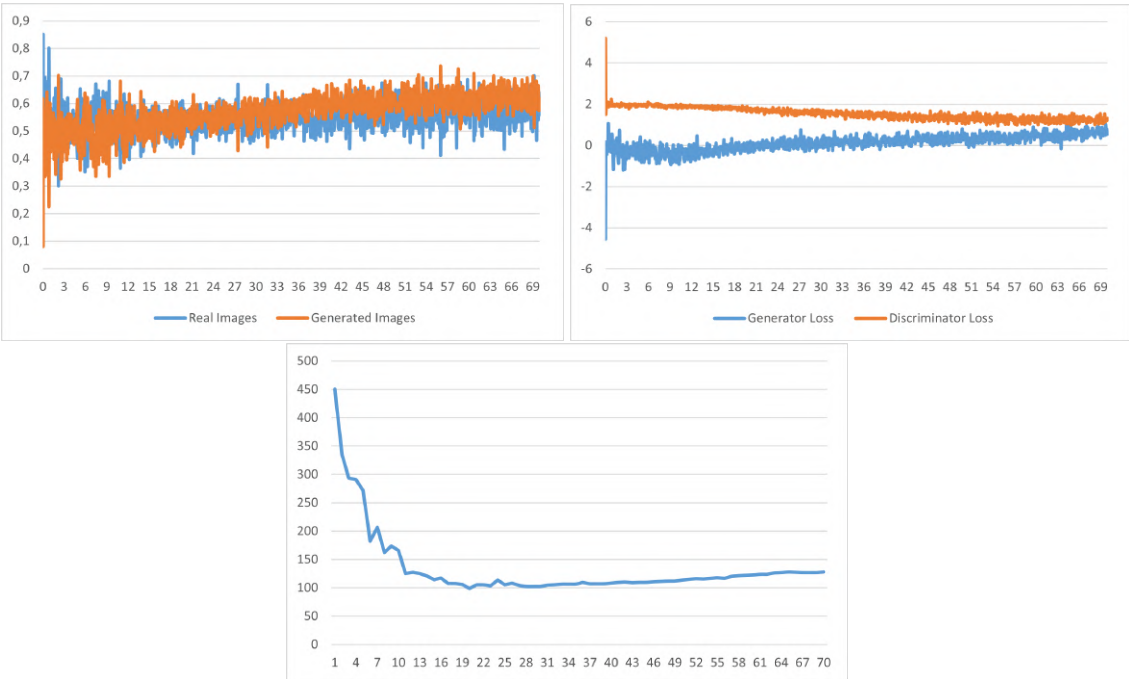


Figure A.18: Discriminator accuracy at predicting the image source (left), loss functions (right) and FID (center) values from the *1000 Image Set* model throughout training

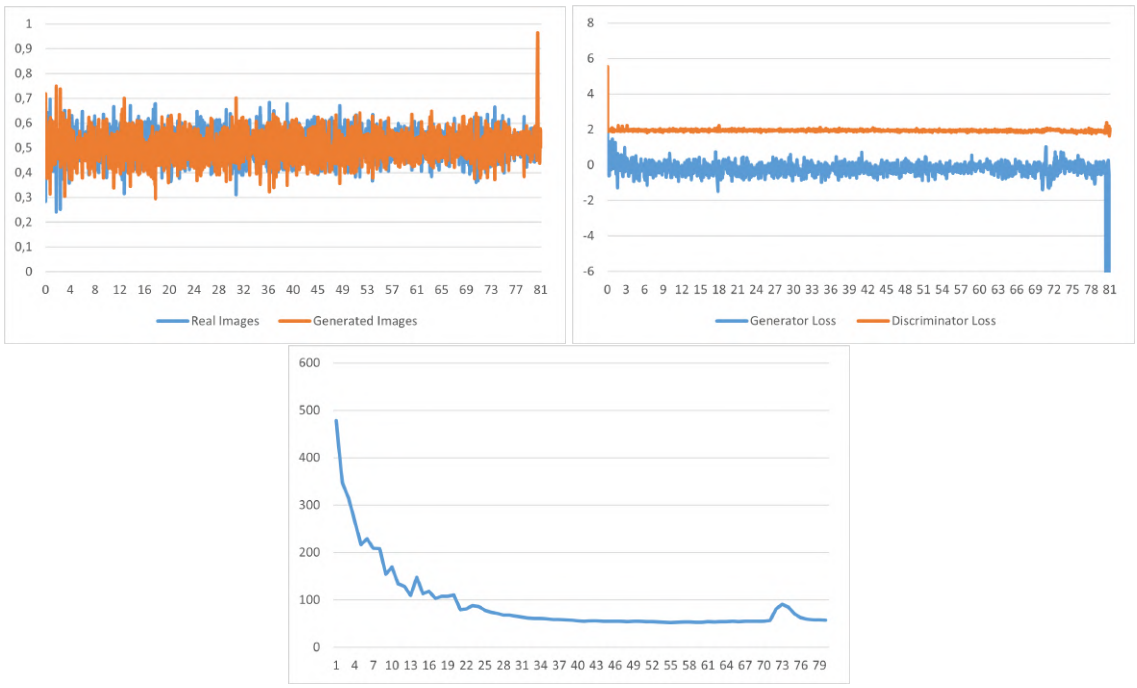


Figure A.19: Discriminator accuracy at predicting the image source (left), loss functions (right) and FID (center) values from the *4000 Image Set* model throughout training

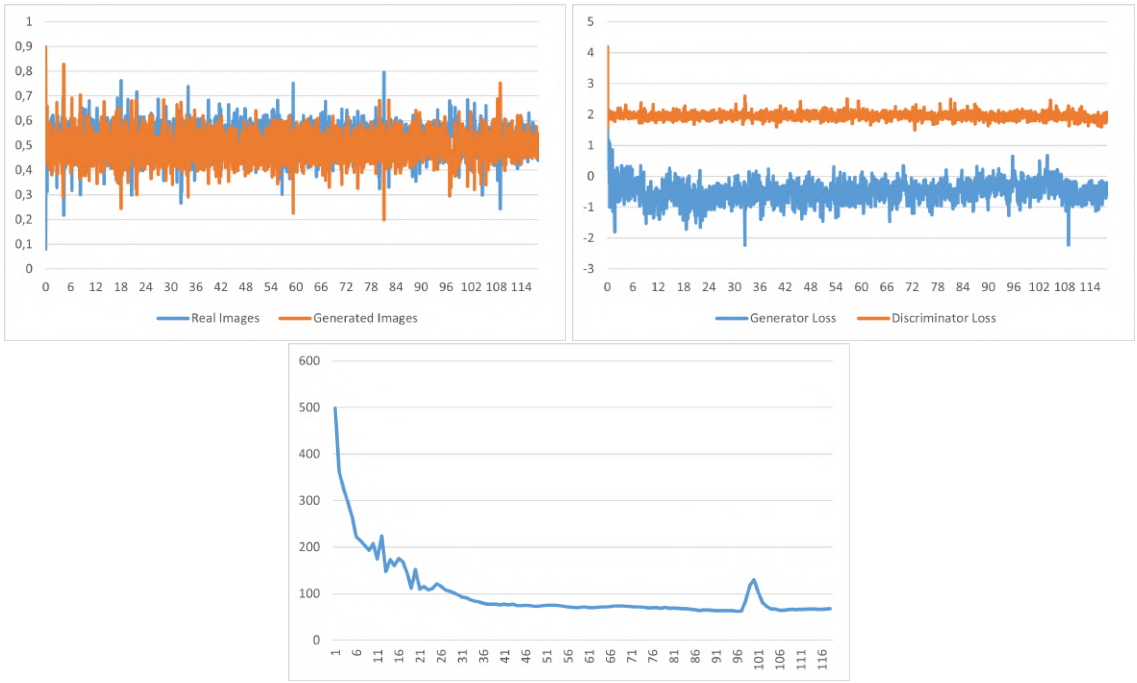


Figure A.20: Discriminator accuracy at predicting the image source (left), loss functions (right) and FID (center) values from the *4000 Image Set - Larger Batch Size* model throughout training

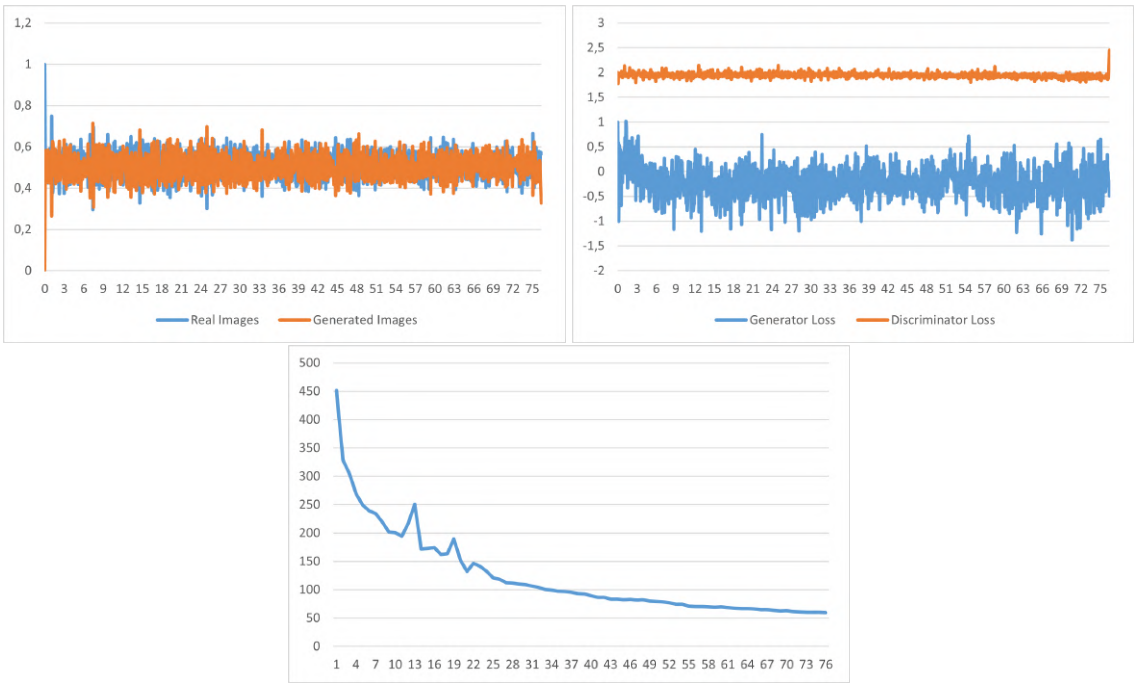


Figure A.21: Discriminator accuracy at predicting the image source (left), loss functions (right) and FID (center) values from the *4000 Image Set - Horizontal Flip* $p = 0.5$ model throughout training